

Using On-line Tutoring Records to Predict End-of-Year Exam Scores: Experience with the ASSISTments Project and MCAS 8th Grade Mathematics

Brian W. Junker*
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213
brian@stat.cmu.edu

December 23, 2006

Abstract

The ASSISTment system is an online benchmark testing system that tutors as it tests. The system has been implemented for the content of the 8th grade Mathematics portion of the Massachusetts Comprehensive Assessment System (MCAS) exams, has been developed and tested in Massachusetts middle schools, and is being adapted for use in other states such as Pennsylvania. Two main 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. In this chapter we focus on the first goal and consider 10 prediction models: how they reflect different models for student proficiency, how they account for student learning over time, and how well they predict MCAS scores. We conclude that a combination of measures, including response accuracy (right/wrong) measures that account for problem difficulty, response efficiency, and help-seeking behavior, produce the best prediction models. In addition, our investigations of prediction models reveal patterns of learning over time that should be captured in feedback reports for teachers.

*This project involves the efforts of many, including principal investigators Neil Heffernan (Worcester Polytechnic Institute) and Ken Koedinger (Carnegie Mellon) as well as Nathaniel O. Anozie, 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 made possible with funding from the US Department of Education, National Science Foundation (NSF), Office of Naval Research, Spencer Foundation, and the US Army.

1 The ASSISTments Project

In many States there are concerns about poor student performance on new high-stakes standards-based tests that are required by United States Public Law 107-110 (the *No Child Left Behind Act* of 2001, NCLB). For instance the Massachusetts Comprehensive Assessment System (MCAS), administers rigorous tests in English, math, history and science in grades 3–12. Students need to pass the math and English portions of the 10th grade versions in order to get a high school diploma without further remediation. In 2003 a full 10% of high school seniors were predicted to be denied a high school diploma due to having failed to pass the test on their fourth try. The problem is most acute with minority students; the failure rates for blacks and Latinos are 25% and 30%, respectively. This problem was viewed with such seriousness that the governor of Massachusetts proposed giving out \$1,000 vouchers to students to get individualized tutoring¹. While that proposal did not get enacted, the Massachusetts Legislature, in a very difficult budget year, increased spending on MCAS extra-help programs by 25% to \$50 million. Moreover, the State of Massachusetts has singled out student performance on the 8th grade math test as an area of highest need for improvement². This test covers middle school algebra, but not the formal algebra (e.g., factoring polynomials) typically done in 9th grade.

Partly in response to this pressure, and partly because teachers, parents, and other stakeholders want and need more immediate feedback about how students are doing, there has recently been intense interest in using periodic benchmark tests to predict student performance on end-of-year accountability assessments (Olson, 2005). Some teachers make extensive use of practice tests and released items to target specific student knowledge needs and identify learning opportunities for individual students and the class as a whole. However, such formative assessments not only require great effort and dedication, but they also take valuable time away from instruction. On-line testing systems that automatically grade students and provide reports (e.g., Renaissance Learning³ or

¹<http://www.edweek.org/ew/newstory.cfm?slug=02mcas.h21>

²<http://www.doe.mass.edu/mcas/2002/results/summary.pdf>

³www.renlearn.com

Measured Progress⁴) reduce the demands on the teacher, however, they do not fundamentally address the formative assessment dilemma: although such assessments intrude on instructional time, they still may be uninformative because they are not based on a sufficiently fine grained model of the knowledge involved, or a sufficiently rich data record for each student.

Another application of technology that has an established record of success in supporting classroom instruction is that of computer based, intelligent tutoring systems. For example, Cognitive Tutors developed at Carnegie Mellon University (e.g., Corbett, Koedinger & Hadley, 2001) combine cognitive science theory, human-computer interaction (HCI) methods, and particular artificial intelligence (AI) algorithms for modeling student thinking. Cognitive Tutors based courses in Algebra, Geometry, and four other areas of high school (e.g., Alevan & Koedinger, 2002) and middle school (e.g., Koedinger, 2002) mathematics have been developed. Classroom evaluations of the Cognitive Tutor Algebra course, for example, have demonstrated 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 (Koedinger et al., 1997; Koedinger, Corbett, Ritter, & Shapiro, 2000).


The ASSISTments⁵ Project (<http://www.assistment.org>) is an attempt to blend the positive features of both computer-based tutoring and benchmark testing. Like most computer-based tutoring systems, the ASSISTment system guides students through the performance of educationally relevant tasks, in this case solving 8th grade mathematics problems. The ASSISTment system also monitors various aspects of students' performance, including speed, accuracy, attempting, and hinting metrics, on which to base prediction of proficiency on the MCAS 8th grade mathematics examination, as well as individual and group progress reports to teachers and others stakeholders, at daily, weekly or other time intervals. Although inspired by needs of Massachusetts students, the ASSISTments System is also being adapted for use in Pennsylvania and potentially other States.

A typical student interaction in the ASSISTments System is built around a single released

⁴www.measuredprogress.org

⁵Coined by Ken Koedinger, to combine the *assisting* and *assessment* functions of the system.

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.

MCAS item, or a morph⁶ 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 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. All questions are coded by source (e.g. MCAS released item, morph of a released item, etc.), and knowledge components (KC’s; e.g. skills, pieces of knowledge, and other cognitive attributes)

⁶In other contexts, e.g. Embretson (1999), item morphs are called “item clones”.

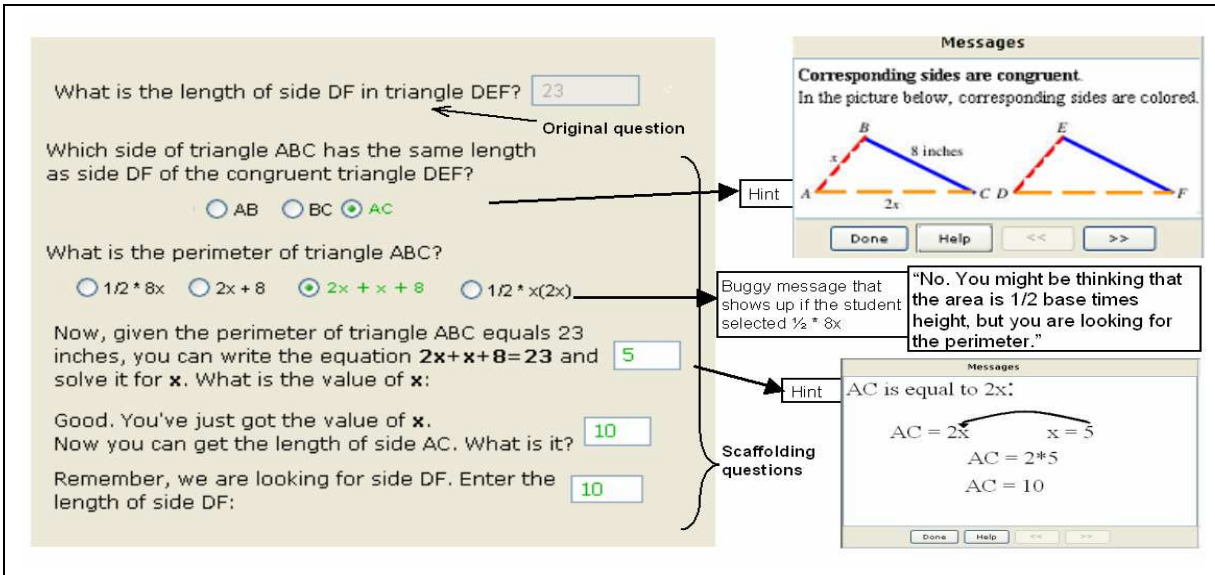


Figure 2: Annotated student interaction with the ASSISTment system, based on the main question in Figure 1.

required. In addition teachers working with project researchers may collect ASSISTment items together into “curricula” or groups of items that are administered to just his or her students. The system tracks individual students through time, recording speed, accuracy and other data, and provides regular reports to teachers per student, per class, etc.

The ASSISTment system is implemented in a more general, extensible tutoring architecture (Razzaq et al., to appear). The architecture is designed to be scalable from simple pseudo-tutors with few users to model-tracing tutors and thousands of users; the ASSISTment system itself is on the simpler end of this range. The architecture consists of

- A Curriculum Unit that allows items to be organized into multiple overlapping curricula, and allows sections within a curriculum to be administered according to a variety of rules, including linear, random, and designed-experiment assignment;
- Problem and Tutoring Strategy Units that manage task organization and user interaction (e.g. main questions and scaffolds, interface widgets, etc.) and allow mapping of task components

(questions) to multiple transfer models⁷; and

- A Logging Unit that provides a fine-grained trace of human-computer interactions as well as various mechanisms for abstracting or coarsening this trace into usable data.

The architecture is supported by a web-based item builder that is used by both research staff and classroom teachers to develop item content, and that provides support for building item curricula, mapping tasks to transfer models, etc. A back-end relational database and networking architecture supports user reports for students, teachers, coaches, administrators, etc., as well as research data analyses.

As indicated above, two main 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 example, 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.). Thus the prediction problem is analogous to shooting at a barn in the fog (students' eventual MCAS scores) from a moving train (students' interactions with the ASSISTment System as they learn throughout the school year).

In addition, different transfer models 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), 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

⁷A *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).

granularity expected by each stakeholder. Third, 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 model⁸.

In this chapter we focus mainly on the task of predicting end-of-year MCAS scores from student-ASSISTment interaction data collected periodically throughout the school year, based on each of the transfer models indicated above, and a variety of psychometric and prediction models. In Section 2 we consider “static” prediction models, in which data is aggregated for some time, usually from September to the time of the MCAS exam. Since they use the most data, these models ought to do the best at prediction of MCAS scores. In Section 3 we consider “dynamic” prediction models, in which predictions are made periodically throughout the school year, based on the data at hand at the time of the prediction.

The data we consider comes 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 by approximately 700 students (each study surveyed below uses a slightly different sample size depending on its goals). The remaining questions and students represented various experimental or otherwise non-usable data for the studies considered here. 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 AS-

⁸It 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.

SISTment items seen by each student varied widely from student to student. Finally, in Section 5 we discuss successes and challenges for ASSISTment-based data collection, prediction, reporting and statistical analysis.

2 Static Prediction Models

Much work has been done in the past 10 years or so on developing “online testing metrics” for dynamic testing (Campione, Brown & Bryant, 1985; Grigorenko & Sternberg, 1998) to supplement accuracy data (wrong/right scores) from a single sitting, in characterizing student proficiency. For example, Table 1 defines a set of such metrics collected by or derivable from the Logging Unit in the ASSISTment System.

In the work described here the goal is to train a prediction function to provide accurate predictions of MCAS scores from such ASSISTment metrics, using 2004–2005 data for which both ASSISTment and MCAS score data are available. In this section we consider “static” predictions, that is, predictions based on ASSISTment data aggregated up to a fixed point in time, usually from September to the time of the MCAS exam. In Section 3 we consider “dynamic” predictions, which are intended to account for or uncover student growth in various ways, and are designed to be used frequently throughout the school year.

The prediction functions we build using the 2004–2005 data are also intended to work well in future years, and so a natural criterion with which to compare candidate prediction functions is cross-validated prediction error. For reasons of interpretability, the prediction error function chosen was mean absolute deviation (MAD),

$$MAD = \frac{1}{n} \sum_{i=1}^n |MCAS_i - pred_i|, \quad (1)$$

where $MCAS_i$ is the actual 2005 MCAS score of the i^{th} student, and $pred_i$ is the predicted score from the prediction function being evaluated. In most cases we also compute mean squared error, MSE (squaring the deviations in the sum in (1)), and root mean squared error, $RMSE = \sqrt{MSE}$.

Table 1: Online testing metrics considered by Anozie and Junker (2006). A similar set of metrics is used by Feng, Heffernan & Koedinger (2006; in press) and Ting & Lee (2005).

Summary Per Month	Description
NumAllMain	Number of complete main questions
NumAllScaff	Number of complete scaffolds
NumCorMain	Number of correct main questions
NumHintsAll	Number of hints on main questions and scaffolds
NumAttAll	Number of attempts
NumSecAll	Number of seconds on main questions and scaffolds
AttCorMain	Number of a attempts on correct main questions
AttIncMain	Number of attempts on incorrect main questions.
AttCorScaf	Number of attempts on correct scaffolds
AttIncScaf	Number of attempts on incorrect scaffolds
SecCorMain	Number of seconds on correct main questions
SecIncMain	Number of seconds on incorrect main questions.
SecCorScaf	Number of seconds on correct scaffolds
SecIncScaf	Number of seconds on incorrect scaffolds
NumCorScaf	Number of correct scaffolds
MedSecAllMain	Median number of seconds on main questions
MedSecIncMain	Median number of seconds on incorrect main questions
PctSecIncMain	percent of time on main questions spent on incorrect main questions
PctCorScaf	percent of scaffolds correct
PctCorMain	Percent of main questions correct
NumPmAllScaf	Number of complete scaffolds per minute
NumPmAllMain	Number of complete main questions per minute
NumIncMain	Number of incorrect main questions
NumIncScaf	Number of incorrect scaffolds
PctSecIncScaf	Percent of time on scaffolds spent on incorrect scaffolds
NumHintsIncMain	Hints plus incorrect main questions
NumHintsIncMainPerMain	Hints plus incorrect main question per ASSISTment

The MCAS score used in (1) is the raw number-right score, which ranges from 0 to 54 in most cases, rather than the scaled reporting score, which ranges from 200 to 280. The MCAS reporting scale is created anew each year by: (a) running a standard-setting procedure to determine achievement levels in terms of raw number-right; and (b) developing a piecewise linear function to transform raw number-right to the 200–280 scale, such that the cutpoints for each achievement level are the same numerical values from year to year (Rothman, 2001). Because of this complication, all of our procedures are judged on their ability to predict the raw number-right score. As additional years’ data are collected, we will compare prediction of raw number-right with prediction of the moving reporting scale target, to see whether the standard setting procedure provides a more stable target from year to year than the raw number-right score. Some analyses do not use all available MCAS questions, and so prediction error will also be reported as a percent of the maximum possible raw score,

$$Pct_Err = MAD/(Max\ Raw\ Score) \quad (2)$$

where “*Max Raw Score*” is the maximum raw score possible with the MCAS questions used (54 points if all 39 MCAS questions are used, since some are scored wrong/right and some are scored with partial credit).

Analyzing 2003–2004 pilot data for the ASSISTment system, Ting & Lee (2005) concluded that such metrics may not contribute much above proportion correct on a paper and pencil benchmark test in predicting end-of-year MCAS scores. However the pilot data set was small (in both number of students and number of items) and the system was very much under development during the pilot data collection period.

Feng, Heffernan & Koedinger (2006; to appear) considered a similar set of online metrics aggregated over seven months from the 2004–2005 school year, for a subset of 600 of the 2004–2005 students. They compared predicting the 54-point raw MCAS score with these summaries, vs. using only paper and pencil tests given before and after these seven months’ use of ASSISTments. Using stepwise variable selection they found the variables listed in Table 2 to be the best predictors:

Table 2: Final stepwise regression model of Feng, Heffernan & Koedinger (2006; to appear).

Predictor	Coefficient
(Const)	26.04
Pre_Test	0.64
Pct_Correct_All	24.21
Avg_Attempts	-10.56
Avg_Hint_Reqs	-2.28

September pre-test score, percent correct on the first attempt at all main questions and scaffolds, average number of attempts per question, and average number of hint requests per question. Although September pre-test score is in the model, it contributes relatively little to the prediction of the raw 54-point MCAS score; instead approximately half the MCAS score is predicted by students' proportion correct on ASSISTment main questions and scaffolds, with substantial debits for students who make many wrong attempts or ask for many hints. This model had within-sample $MAD = 5.533$ and $Pct_Err = MAD/54 = 10.25\%$. These error rates are lower bounds on cross-validation error rates.

Ayers & Junker (2006) improved on the Feng, Heffernan & Koedinger (2006; to appear) approach, for a subset of 683 students, by replacing percent correct with an item response theory (IRT; van der Linden & Hambleton, 1997) score based on main questions only, to account for the varying difficulty of the different samples of questions that each student sees. They considered both a generic Rasch (1960/1980) model, in which the probability of a correct response X_{ij} for student i on main question j is modeled as a logistic regression⁹

$$\text{logit } P[X_{ij} = 1 \mid \theta_i, \beta_j] = \theta_i - \beta_j \quad (3)$$

depending on student proficiency θ_i and question difficulty β_j ; and a linear logistic test model (LLTM; Fischer & Molenaar, 1995) in which main question difficulty was decomposed into com-

⁹Recall that $\text{logit } p = \ln p/(1 - p)$, so that if $\text{logit } p = \lambda$, then $p = e^\lambda/(1 + e^\lambda)$.

ponents for each KC needed to answer the main question,

$$\text{logit } P[X_{ij} = 1 \mid \theta_i, \alpha_1, \dots, \alpha_K] = \theta_i - \sum_{k=1}^K Q_{kj} \alpha_k \quad (4)$$

where $Q_{kj} = 1$ if KC k contributes to the difficulty of question j , and 0 otherwise, and α_k is the contribution of KC k to the difficulty of any question involving that KC. Only 77 KC's from the WPI-106 transfer model (Pardos, et al., 2006) were needed to model the 354 main questions they considered, so the Q -matrix here is 77×354 . There is a bias-variance tradeoff question here for prediction: the unrestricted Rasch model will produce less-biased proficiency estimates on which to base prediction of MCAS scores; but the LLTM with hundreds fewer parameters will produce lower-variance proficiency estimates.

They found that the Rasch model fit dramatically better (reduction in BIC¹⁰ of 3,300 for an increase of 277 parameters), so the lower bias of the Rasch model should win over the lower variance of the LLTM for prediction. As shown in Figure 3, there is some evidence that a moderately finer-grained transfer model might have worked better with the LLTM. For example, it is especially clear that several 1-KC main questions, that depend on the same KC and hence have the same difficulty estimate under the LLTM, have greatly varying difficulty estimates under the unconstrained Rasch model. This suggests that sources of difficulty not accounted for by the WPI-106 transfer model are present in those problems.

Ayers & Junker (2006) then incorporated the student proficiency score θ_i for each student i from the Rasch model into a prediction model for raw MCAS scores of the form

$$MCAS_i = \lambda_0 + \lambda_1 \theta_i + \sum_{m=2}^M \lambda_m Y_{im} + \varepsilon_i \quad (5)$$

where $MCAS_i$ is the student's raw MCAS score, Y_{im} are values of online testing metrics selected from Table 1, ε_i is residual error, and the regression coefficients λ_m , $m = 0, \dots, M$ are to be estimated. In order to account for measurement error in θ_i , the model (5) is estimated using the

¹⁰Bayes Information Criterion, also known as the Schwarz Criterion; see for example Kass & Raftery (1995, p. 778).

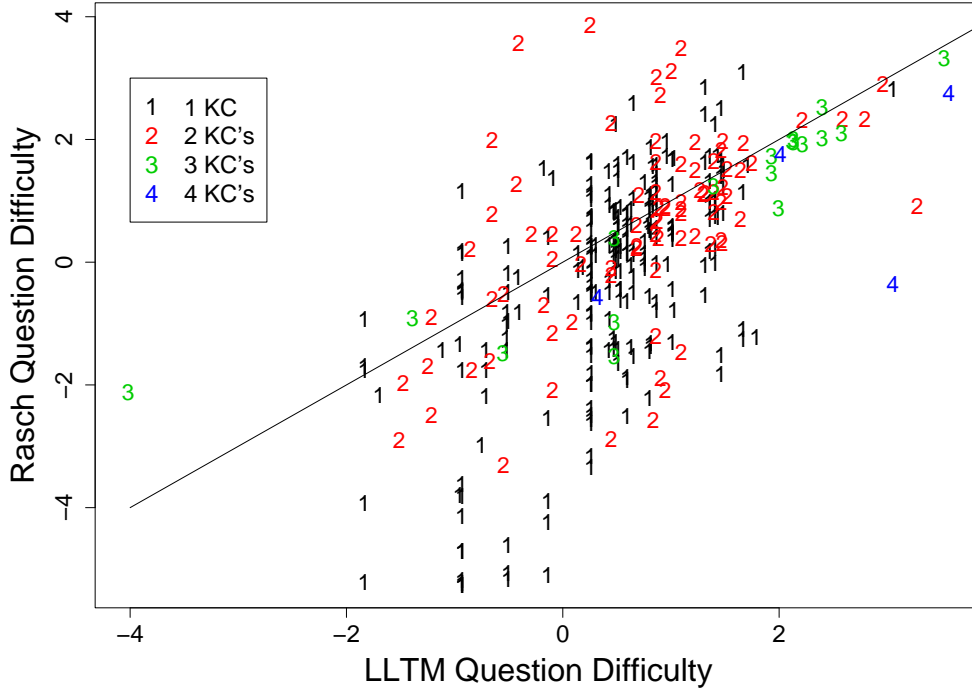


Figure 3: Comparing main question difficulty estimates from the LLTM (horizontal axis) with main question difficulty estimates from the Rasch model (Ayers & Junker, 2006). The number of KC's from the WPI-106 required for each question is also indicated.

WinBUGS (Spiegelhalter, Thomas, & Best, 2003) software, in which multiple imputations, or plausible values (e.g. Mislevy, 1991), were generated for θ_i from the fitted Rasch model, using a Markov Chain Monte Carlo (MCMC) algorithm. (This model was introduced by Schofield, Taylor & Junker, 2006, to analyze the influence of literacy on income using the National Adult Literacy Survey.)

They evaluated their results using 10-fold cross-validated *MAD* and *Pct_Err* for predicting the 54-point raw MCAS score. For a model using only percent-correct on main questions (and neither Rasch proficiency scores nor online metrics), they found $CV-MAD = 7.18$ ($CV-Pct_Err = 13.33\%$). Replacing percent-correct with Rasch proficiency calculated from main questions only,

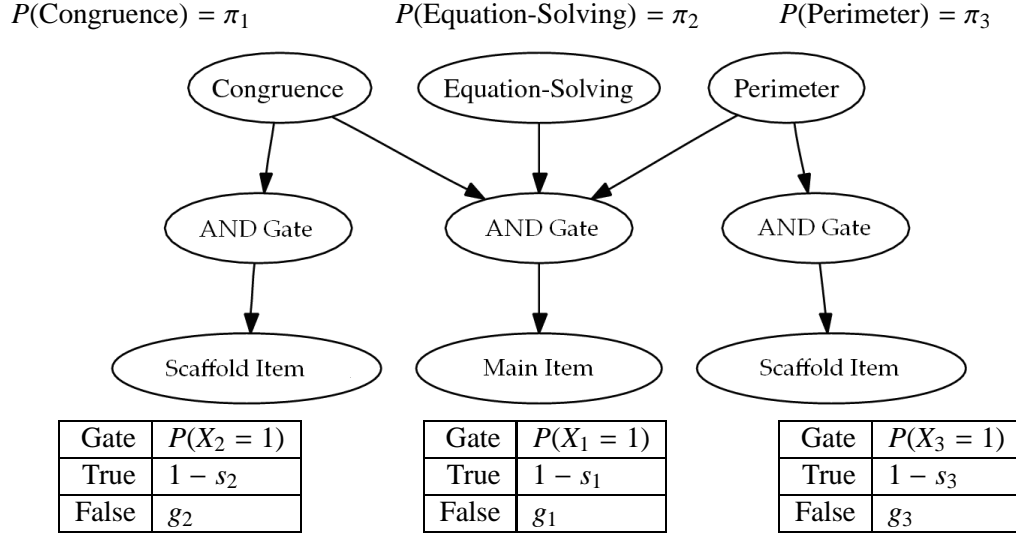


Figure 4: Illustration of the conjunctive Bayes Net (DINA) model, after Pardos et al. (2006). π_k is the base rate (prior probability) of KC k in the student population; and for each question X_j , s_j is the probability of a slip leading to an incorrect answer, and g_j the probability of a guess leading to a correct answer.

they found $CV-MAD = 5.90$ ($CV-Pct_Err = 10.93\%$), obtaining a cross-validation result comparable to Feng, Heffernan & Koedinger (2006; to appear) within-sample results by replacing their online metrics and pretest scores with only the Rasch proficiency estimate. Finally, combining Rasch proficiency with five online metrics from Table 1 chosen to minimize MAD in a forward selection scheme they found $CV-MAD = 5.24$ ($CV-Pct_Err = 9.70\%$), improving on the earlier within-sample results. All of the online metrics included in their final model were related to the efficiency of student work (NumPmAllScaf; see Table 1 for definition) or contrasts between the time spent answering correctly or incorrectly (SecCorScaff, SecIncMain, MedSecIncMain, and PctSecIncMain; see Table 1 for definitions).

Pardos et al. (2006) and Anozie (2006) considered static prediction using conjunctive Bayes Nets (Maris, 1999; Mislevy, Almond, Yan & Steinberg, 1999; Junker & Sijtsma, 2001) for binary KC's (1 = learned, 0 = unlearned) & responses (1 = correct, 0 = incorrect). A schematic illustration

of the model, also called the “deterministic input, noisy AND-gate” (DINA) model by Junker & Sijtsma (2001), is presented in Figure 4 for a main question (X_1) and two scaffolds (X_2 and X_3). In the figure, the main question depends on three KC’s, “Congruence”, “Equation Solving” and “Perimeter”. The two scaffold questions focus on “Congruence” and “Perimeter” respectively. Each KC k has a population base rate (prior probability of already being known to the student) of π_k . The probability of student i getting a correct answer on question j (whether it is a main question or a scaffold) is expressed in terms of a guessing parameter g_j and a slip parameter s_j for that question,

$$P[X_{ij} = 1 \mid g_j, s_j] = \begin{cases} (1 - s_j), & \text{if student } i \text{ knows all the KC's} \\ & \text{relevant to question } j; \\ g_j, & \text{if not.} \end{cases} \quad (6)$$

The mapping of KC’s relevant to each question is accomplished with a Q -matrix, as in the LLTM. A key difference in the models is that KC’s combine additively to determine question difficulty in the LLTM, whereas they combine conjunctively to determine cognitive demand in the Bayes Net model.

Pardos et al. (2006) compared the predictive accuracy of conjunctive Bayes Nets based on several different transfer models, for a 30-item, 30-point subset of the MCAS: a one-binary-KC model, a five-binary-KC model corresponding to the five MCAS strands, a 39-binary-KC model corresponding to the 39 MCAS learning standards, and a 106-binary-KC model based on the WPI-106 transfer model. They fixed the guessing parameters $g_j \equiv 0.10$ and slip parameters $s_j = 0.05$ for all items, fixed the base-rate probabilities $\pi_k = 0.5$ for all KC’s, inferred which KC’s each student had learned, based on seven months’ data using the Bayes Net Toolbox (Murphy, 2001), and predicted success on individual MCAS questions by mapping KC’s from the transfer model to the released MCAS questions. In this analysis the most successful model was the 39-KC model, with $MAD = 4.5$ and $Pct_Err = MAD/30 = 15.00\%$.

Anozie (2006) focused on subsets of the Ayers & Junker (2006) data from the first three months of the 2004–2005 data, involving 295 students and approximately 300 questions tapping a 62-KC

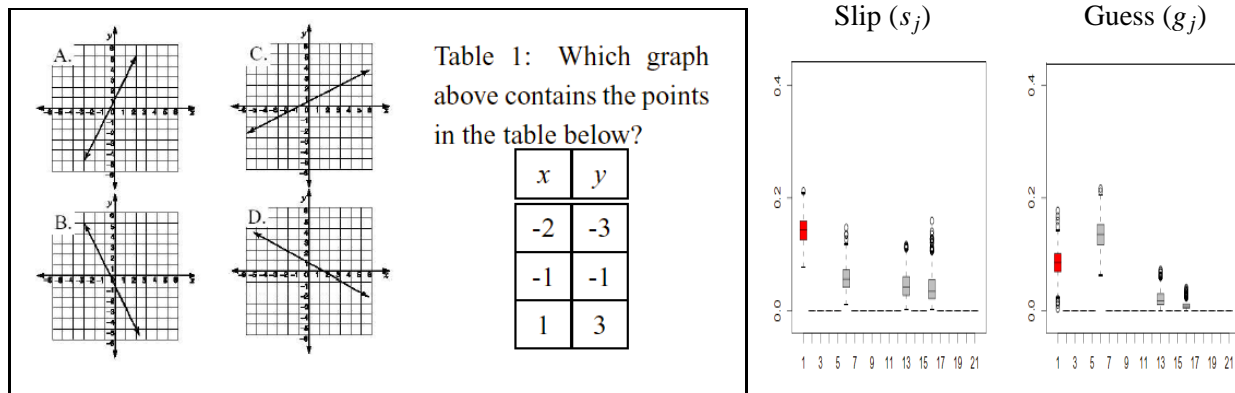


Figure 5: An ASSISTment main question and estimates of its slip (s_j) and guessing (g_j) parameters (red boxes; the grey boxes are for other main questions). Anozie (2006).

subset of the WPI-106 transfer model, and estimated g_j 's, s_j 's and π_k 's from the data, using an MCMC procedure developed for the statistical package R (The R Foundation, 2006). The raw 54-point MCAS score was predicted as a linear function of the raw number of KC's learned according to the Bayes Net model, for each student. 10-fold cross-validation of prediction using two months' ASSISTment data yielded $CV-MAD = 8.11$, and $CV-Pct_Err = 15.02\%$. When three months' ASSISTment data were used, $CV-MAD$ and $CV-Pct_Err$ were reduced to 6.79 and 12.58, respectively.

Although the predictive error results were disappointing compared to the simpler models of Feng, Heffernan & Koedinger (2006; to appear) and Ayers & Junker (2006), the Bayes Net models yield diagnostic information that may be of interest to teachers, quite apart from predicting MCAS scores. Close analysis is also revealing about the WPI-106 as a KC measurement model.

Consider, for example, Figures 5 and 6. The left part of Figure 5 shows a main question tagged with a single KC, "Plotting Points", by the WPI-106 transfer model. On the right in Figure 5 are summaries of the estimated slip and guess parameters for this main question (the middle line in the red box plot is the estimate; the box and whiskers show the extent of uncertainty about the estimate; the grey boxes are for other main questions tagged with other combinations of KC's).

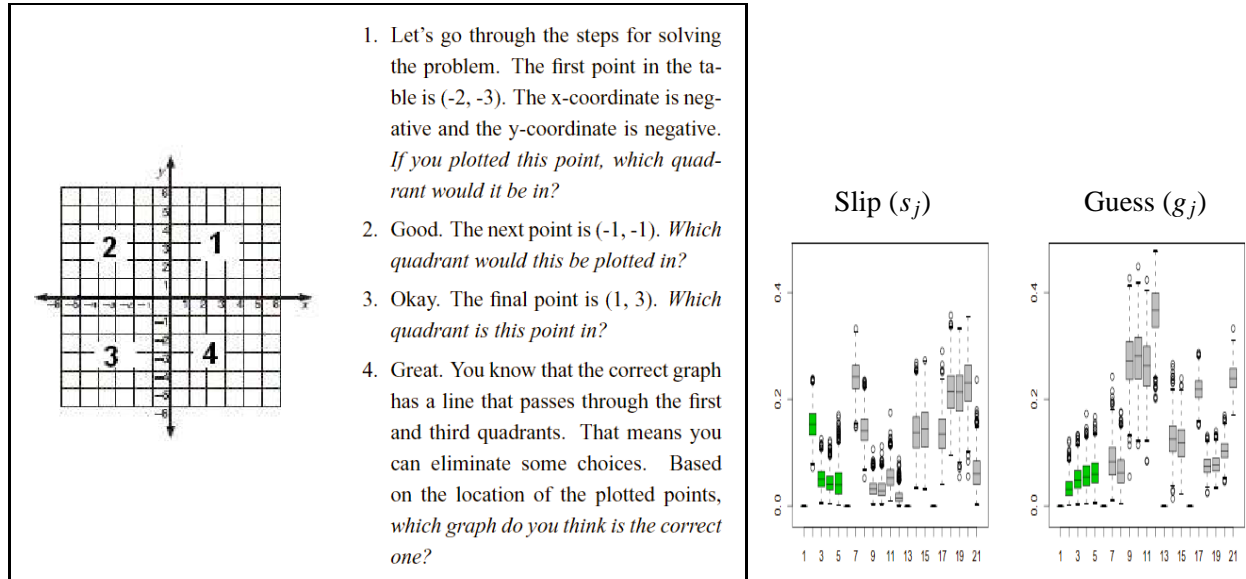


Figure 6: The scaffold questions corresponding to the main question in Figure 5, and estimates of their slip (s_j) and guessing (g_j) parameters (green boxes; the grey boxes are for other scaffold questions). Anozie (2006).

The left part of Figure 6 shows the corresponding scaffolding questions (again each tagged with the same KC, “Plotting Points”) and their estimated slip and guess parameters are shown on the right (green boxes; the grey boxes are for other scaffold questions tagged with other KC’s). 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 Figure 6 as well: 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 decision about the “completeness”, to use DiBello et al.’s (1995) term, of the transfer model: there is a tradeoff to make

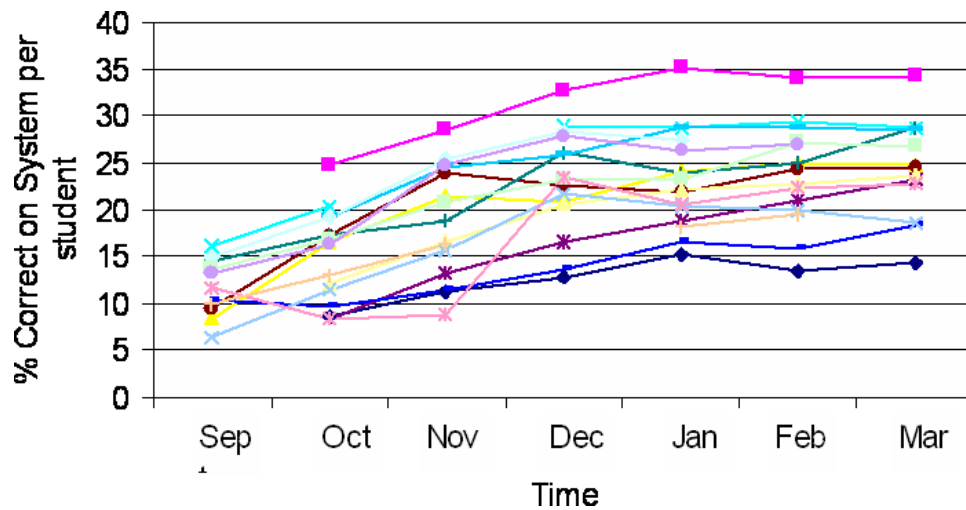


Figure 7: Average percent-correct on ASSISTment main questions for each classroom (colored lines) in the 2004–2005 ASSISTments study, by month. From Razzaq et al. (2005).

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).

3 Dynamic Prediction Models

Figure 7 displays the percent correct on ASSISTment main questions in each month of the 2004–2005 school year, for each class (colored lines) participating in the ASSISTments study. It is clear from the figure that the ASSISTment system is sensitive to student learning, and that students on the whole are improving as the school year progresses. Some of this student learning is due the experiences students are having in school outside the ASSISTment system, and some is due to the ASSISTment system itself (Razzaq et al., 2005). What is less clear is how best to account for this learning in predicting MCAS scores at the end of the year.

Feng, Heffernan & Koedinger (to appear) addressed the heterogeneity in Figure 7 directly by

building growth-curve models in an HLM (hierarchical linear model; see for example Singer & Willett, 2003) framework. They first examined an overall growth-curve model of the form

$$\begin{aligned}\text{Level 1: } \bar{X}_{ti} &= \beta_0 + \beta_1(\text{Month}_t) + \varepsilon_{ti} \\ \text{Level 2: } \beta_0 &= \beta_{00} + \beta_{01}(\text{Covariate}_{0i}) + \varepsilon_{0i} \\ \beta_1 &= \beta_{10} + \beta_{11}(\text{Covariate}_{1i}) + \varepsilon_{1i}\end{aligned}$$

where \bar{X}_{ti} is percent-correct on ASSISTment main questions for student i in month t , Month_t is the number of months into the study (0 for September, 1 for October, etc.), and $\text{Covariate}_{\ell i}$ is an appropriate Level 2 covariate for β_ℓ . For Level 2 covariates, they compared School, Class and Teacher using the BIC measure of fit, and found School to be the best Level 2 covariate for baseline achievement (β_0) and rate of change (β_1), suggesting that School demographics dominate the intercept and the slope in the model.

They also explored differential learning rates for the five MCAS strands (Number & Operations, Algebra, Geometry, Measurement, Data Analysis & Probability) by considering a growth-curve HLM of the form

$$\begin{aligned}\text{Level 1: } \bar{X}_{sti} &= \beta_0 + \beta_1(\text{Quarter}_t) + \varepsilon_{ti} \\ \text{Level 2: } \beta_0 &= \beta_{00} + \beta_{01}(\text{Covariate}_{0i}) + \beta_{02s} + \varepsilon_{0si} \\ \beta_1 &= \beta_{10} + \beta_{11}(\text{Covariate}_{1i}) + \beta_{12s} + \varepsilon_{1si}\end{aligned}$$

where now \bar{X}_{sti} is the proportion correct on ASSISTment main items in strand s at time t for student i , Quarter_t is the school-year quarter (0, 1, 2, or 3) at time t , and again $\text{Covariate}_{\ell i}$ is an appropriate Level 2 covariate for β_ℓ . Again using the BIC measure of fit, they found that Strand matters for both the baseline level of achievement (β_0) and rate of change (β_1). No other covariates were needed to predict rate of change (β_1), but the September pre-test score was an additional useful predictor for baseline achievement β_0 (thus replacing School in their first model).

After the growth curve model is fitted, one could extrapolate in time to the month of the MCAS exam to make a prediction about the student's MCAS score. Feng, Heffernan, Mani & Heffernan (2006) tried this approach with growth curve models for individual questions. Letting X_{ij} be the

0/1 response of student i on question j tapping KC k in month t , they considered the logistic growth curve model

$$\left. \begin{array}{lcl} \text{Level 1: } \text{logit } P[X_{ijkt} = 1] & = & (\beta_0 + \beta_{0k}) + (\beta_1 + \beta_{1k})(Month_t) \\ \text{Level 2: } & & \beta_0 = \beta_{00} + \varepsilon_{0i} \\ & & \beta_1 = \beta_{10} + \varepsilon_{1i} \end{array} \right\} \quad (7)$$

where again $Month_t$ is elapsed month in the study (September = 0, October = 1, etc.) and β_{0k} and β_{1k} are respective fixed effects for baseline and rate of change in probability of correctly answering a question tapping KC k .

This model is equivalent to an LLTM-style restriction of the Rasch model of equation (3), where now: (a) student proficiency is allowed to depend on time, $\theta_i = (\beta_{00} + \varepsilon_{0i}) + (\beta_{10} + \varepsilon_{1i})(Month_t)$; and (b) question difficulty is allowed to depend on KC and time: $-\beta_j = (\beta_0 + \beta_{0k}) + (\beta_1 + \beta_{1k})(Month_t)$. Rather than implementing a full Q-matrix mapping of multiple KC's onto each question as in the LLTM of equation (4), Feng, Heffernan, Mani & Heffernan (2006) assigned only the most difficult KC in the transfer model for each question (according to lowest proportion correct among all questions depending on each KC) to that question.

They fitted the model in equation (7) using the `lme4` library in R (The R Foundation, 2006), extrapolated the fitted model to the time of the MCAS exam to obtain probabilities of getting each MCAS question correct (using the same max-difficulty reduction of the transfer model for the released MCAS items) and summed these probabilities to predict the students' raw scores for a 34-point subset of MCAS questions. The prediction error rates for this method, using a subset of 497 students who answered an average of 89 main questions and 189 scaffolds depending on 78 of the WPI-106 KC's, were comparable to those of the Bayes Net approach: $MAD = 4.121$, $Pct_Err = MAD/34 = 12.12\%$.

One might try to improve on this prediction error by including demographic and related variables, for example $School_i$, as in the hierarchical model of Feng, Heffernan & Koedinger (to appear). Whether one does so depends on one's goals, for both the transfer model itself, and for the portability of the prediction system. If the transfer model is incomplete—does not adequately

account for the cognitive challenge of answering questions—then demographic variables might reasonably be proxies for presence or absence of KC’s unaccounted for in the transfer model. On the other hand, if the transfer model is relatively complete, then we will not need demographic variables for this purpose, and the prediction function is more likely to generalize to novel demographic situations.

A rather different approach was pursued by Anozie & Junker (2006), who looked at the changing influence of online ASSISTment metrics on MCAS performance over time. They computed monthly summaries of each of the online metrics listed in Table 1, and built several linear prediction models, predicting end-of-year raw MCAS scores for each month, using all the online metric summaries available in that and previous months. Since there were seven months of data, seven regression models were built, e.g.

$$\left. \begin{aligned}
 MCAS_i &= \beta_{01} + \beta_{111}(\text{PctCorMain})_i^{oct} + \beta_{211}(\text{PctCorScaf})_i^{oct} \\
 &\quad + \cdots + \epsilon_i^{oct} \\
 MCAS_i &= \beta_{02} + \beta_{121}(\text{PctCorMain})_i^{oct} + \beta_{221}(\text{PctCorScaf})_i^{oct} \\
 &\quad + \beta_{122}(\text{PctCorMain})_i^{nov} + \beta_{222}(\text{PctCorScaf})_i^{nov} \\
 &\quad + \cdots + \epsilon_i^{nov} \\
 MCAS_i &= \beta_{03} + \beta_{131}(\text{PctCorMain})_i^{oct} + \beta_{231}(\text{PctCorScaf})_i^{oct} \\
 &\quad + \beta_{132}(\text{PctCorMain})_i^{nov} + \beta_{232}(\text{PctCorScaf})_i^{nov} \\
 &\quad + \beta_{133}(\text{PctCorMain})_i^{dec} + \beta_{233}(\text{PctCorScaf})_i^{dec} \\
 &\quad + \cdots + \epsilon_i^{dec} \\
 \vdots &\quad \quad \quad \vdots
 \end{aligned} \right\} \quad (8)$$

where $MCAS_i$ is student i ’s actual raw 54-point MCAS score, $(\text{PctCorMain})_i^{oct}$ is student i ’s percent-correct on main questions in October, $(\text{PctCorScaf})_i^{nov}$ is student i ’s percent-correct on scaffold questions in November, and so forth. All metrics in Table 1, not just PctCorMain and PctCorScaf, were considered in these models.

Anozie & Junker (2006) used the same data set as Ayers & Junker (2006). After imputing miss-

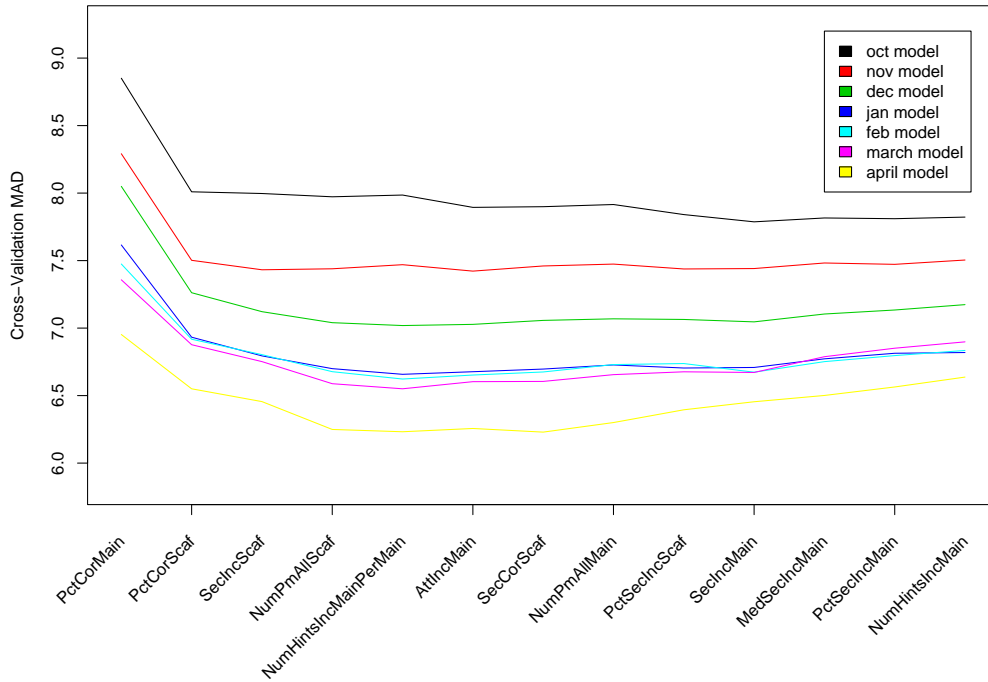


Figure 8: Variable selection for the monthly prediction models of Anozie & Junker (2006). Variables are listed across the horizontal axis according to greedy reduction in *CV-MAD*. Each curve shows the effect on *MAD* of additional variables to the corresponding model in equation (8).

ing summaries (e.g. some students skipped an entire month on the ASSISTment system and their summaries for that month were copied forward from the most current month in which summary data was available), they developed software in R (The R Foundation, 2006) to perform variable selection, using 10-fold cross-validation MAD summed across all seven models. To enhance interpretation, variable selection was done by metric, not by monthly summary, and metrics were included or excluded simultaneously in all seven models: thus if a variable was included, all of its relevant monthly summaries would be included in all seven regression models. By constraining the variable selection in this way, Anozie & Junker (2006) could track the relative influence of adding more metrics, vs. adding more months of summaries, for example.

A summary of this variable selection procedure is presented in Figure 8. Variables are listed across the horizontal axis according to greedy reduction in *CV-MAD*: PctCorMain was most often added first in 100 replications of the cross-validation variable selection procedure; PctCorScaf was most often added second in the same 100 replications, etc. Each curve shows the effect on *MAD* of additional variables to the corresponding model in equation (8): the top curve shows the effect on *MAD* of adding successively more October summaries of variables to the first model in equation (8); the next curve shows the effect on *MAD* of adding successively more October and November summaries of variables to the second model in (8); and so forth.

It is clear from inspection of Figure 8 that adding more months' of data helps in prediction more than adding more online metrics. Most models in the figure have stable or minimum *MAD*'s when five variables are included in the model: two accuracy variables (PctCorMain, PctCorScaf), two time/efficiency variables (SecIncScaf, NumPmAllScaf), and one variable related to help-seeking (NumHintsIncMainPerMain); see Table 1 for variable definitions. The nature of these variables is consistent with the results of Feng, Heffernan & Koedinger (2006; to appear) and Ayers & Junker (2006).

We can also inspect Figure 8 above the location marked by NumHintsIncMainPerMain, to see what the *CV-MAD* and *CV-Pct_Err* are for all seven of the five-variable monthly models. *CV-MAD* ranges from approximately 8.00 ($CV-Pct_Err = MAD/54 = 14.8\%$) for the model based on October summaries only, to about 6.25 ($CV-Pct_Err = MAD/54 = 11.6\%$) for the model based on all seven monthly summaries of each variable. Most of the improvement in prediction error has already occurred by January.

Figure 9 shows the predicted increase in the raw (54-point) MCAS score corresponding to a 10% increase in each monthly summary of PctCorMain (percent correct on main questions), in each of the seven monthly models above. The different models are indicated by vertical bands and the monthly summaries relevant to each model are connected by line segments. Coefficients significantly different from zero ($p \leq 0.05$) are plotted as filled points; non-significant coefficients are unfilled.

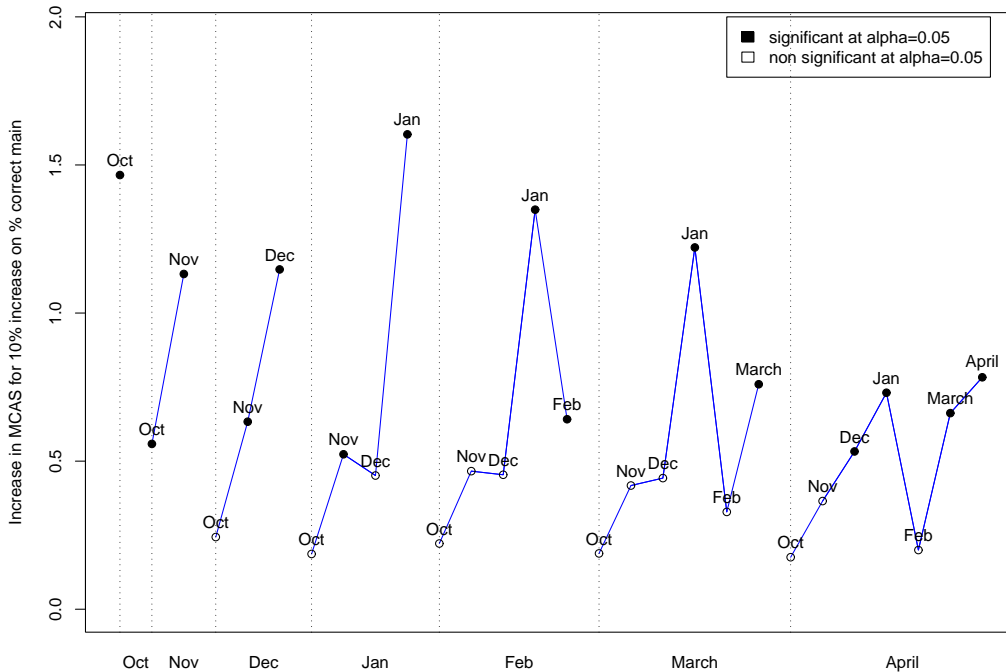


Figure 9: Influence of each monthly summary of PctCorMain (percent correct on main questions) on MCAS prediction (Anozie & Junker, 2006). Different models are indicated by vertical bands, and different monthly summaries within each model are connected by line segments.

Two patterns are clear in Figure 9: first, each monthly summary generally decreases in importance as more recent data are included in the models. For example, the October summary of PctCorMain is a significant predictor only in the October and November models, and the influence that it has on predicting MCAS scores decreases monotonically across the monthly prediction models. Second, within each monthly model, the predicted influence of more recent summaries are generally at least one raw MCAS point higher than the predicted influence of less recent summaries¹¹. This may be another form of evidence of learning (increasing achievement) in the students, as time

¹¹This pattern is less pronounced in later months' models, partly because these models include more monthly summaries as predictors, and the summaries tend to be correlated with one another.

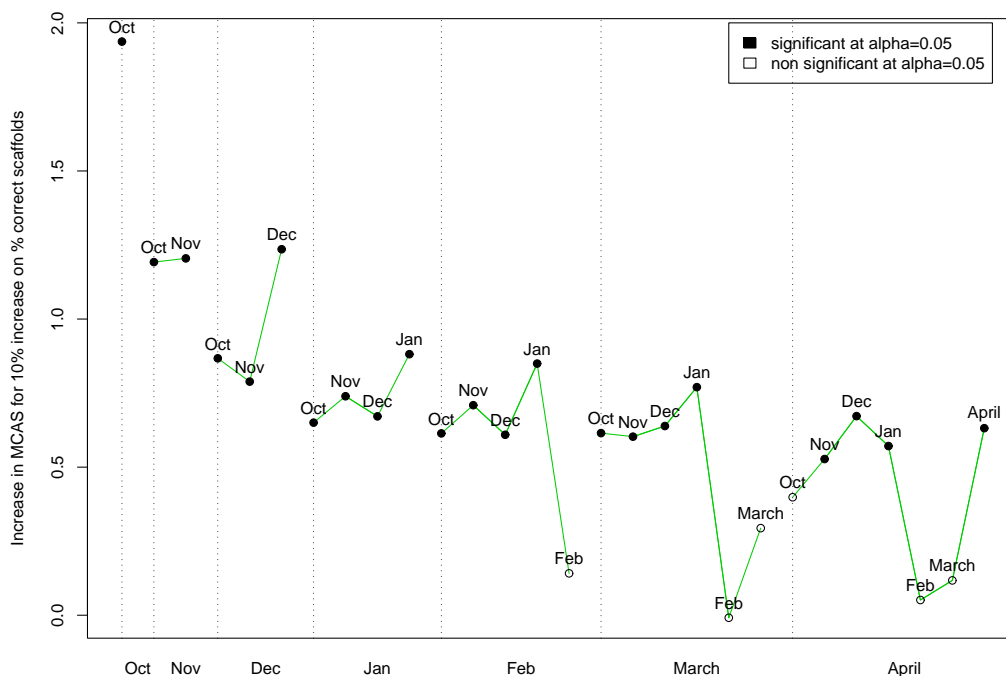


Figure 10: Influence of each monthly summary of PctCorScaf (percent correct on scaffold questions) on MCAS prediction (Anozie & Junker, 2006). Different models are indicated by vertical bands, and different monthly summaries within each model are connected by line segments.

passes during the school year (compare Figure 7).

The exception to this second pattern is the February summary, which appears to have little or no predictive value for MCAS scores. Further investigation revealed that an experiment had been run on the ASSISTment system in February in which a “forced scaffolding” regime was compared to the usual regime in which scaffolds were only provided if the answer to the main question was wrong. To implement the forced scaffolding regime, main questions were scored wrong whether or not the student answered correctly. Thus, many artificially low scores are present in the February data. We believe this is why the February summary does not appear to be useful in the model.

Similarly, Figure 10 shows the predicted increase in the raw (54-point) MCAS score corre-

sponding to a 10% increase in each monthly summary of PctCorScaf (percent correct on scaffold questions). Neither the February nor the March summaries appear to be useful in any model. Although we are not satisfied that we understand the behavior of the March summary, we believe the February anomaly is again due to the “forced scaffolding” experiment in February, since many people who did not need scaffold questions nevertheless were forced to answer them in February, artificially inflating February PctCorScaf scores.

The first pattern seen in Figure 9 is evident in Figure 10: each monthly summary becomes less influential as later data is added to the models. The second pattern, that later summaries are more influential than earlier ones, holds to some extent in the earlier months’ models (October through January). But it is striking that the difference in influence among summaries in the same model is now as little as 0.25 raw MCAS points (compared to a point or more for PctCorMain, as described above). This may reflect the fact that the ASSISTment system generally presents scaffolding questions only when the student is unsure or doesn’t know the material well, so that percent correct on scaffolds reflects learning style or rate, rather than achievement level.

4 Comparing Predictions

A wide variety of strategies have been developed to predict MCAS scores from ASSISTment data. Table 3 summarizes 10 prediction models discussed in this chapter and gives their *MAD* and *Pct_Err* scores. Several conclusions can be drawn from this table.

First, there appears to be a tradeoff between accurately modeling student-to-student variation in proficiency, vs. including online metrics from Table 1. For example, rows 4 and 5 of the table compare two models based on three months’ student data: a model tracking three monthly summaries of five online metrics (Anozie & Junker, 2006, December model), and a 62-KC Bayes net model (Anozie, 2006, December model). The *MAD* prediction errors for the two models are quite comparable, 7.00 and 6.68 respectively. Similarly, rows 8 and 9 compare regression on percent correct on all (main and scaffold) ASSISTment questions, a pretest score, and two online metrics (Feng,

Table 3: Comparison of methods for predicting MCAS scores from ASSISTment data.

Model	Number of Months' Data	Number of Predictors	CV- MAD	Max Raw MCAS Score	CV- Pct_Err
Direct Bayes Net prediction ^(a) (Pardos et al., 2006)	7	39	4.50 ^(a)	30	15.00 ^{(a),(b)}
Regression on PctCorMain + 4 online metrics (Anozie & Junker, 2006, October model)	1	5	8.00	54	14.81
Regression on PctCorMain (Ayers & Junker, 2006)	7	1	7.18	54	13.30
Regression on PctCorMain + 4 online metrics (Anozie & Junker, 2006, December model)	3	15	7.00	54	12.96
Regression on number of KC's learned in Bayes Net (Anozie, 2006, December model)	3	1 ^(c)	6.63	54	12.58
Logistic Growth Curve Model for Questions (Feng, Heffernan, Mani & Heffernan, 2006)	7	78	4.21 ^(b)	34	12.12 ^(b)
Regression on PctCorMain + 4 online metrics (Anozie & Junker, 2006, April model)	7	35	6.25	54	11.57
Regression on PctCorAll, Pretest + two online metrics (Feng, Heffernan & Koedinger, 2006)	7	4	5.53 ^(b)	54	10.25 ^(b)
Regression on Rasch proficiency (Ayers & Junker, 2006)	7	1 ^(d)	5.90	54	10.93
Regression on Rasch proficiency + 5 online metrics (Ayers & Junker, 2006)	7	6 ^(d)	5.24	54	9.70

^(a) Within-sample, not cross-validated.

^(b) Using fixed $g_j = 0.10$, $s_j = 0.05$ and $\pi_k = 0.50$. Subsequently, Pardos, Feng, Heffernan & Heffernan (2006) showed that approximately 3 percentage points of this *Pct_Err* is attributable to prediction bias due to lack of model fit.

^(c) Number of KC's was estimated after fitting 300-item, 62-KC DINA model using MCMC (approx. 600 parameters).

^(d) Proficiencies were estimated after fitting 354-item Rasch model using MCMC (approx. 355 parameters).

Heffernan & Koedinger, 2006), with regression on student proficiency computed from the Rasch model (Ayers & Junker, 2006). The *MAD* scores are again comparable, 5.53 and 5.90 respectively.

Second, greater complexity in the student proficiency model can be helpful, as pointed out in the detailed analyses reported in Pardos et al. (2006) and in Feng, Heffernan, Mani & Heffernan (2006); however a simpler proficiency model that accurately accounts for ASSISTment question difficulty, such as the Rasch model fitted by Ayers & Junker (2006), can substantially improve prediction error. In addition, combining a good proficiency model with suitable online metrics produces the best prediction model.

Third, while the various methods have ultimately produced improvements in prediction error, it seems difficult to get the error below approximately 10% of the maximum possible raw MCAS score. There is, in fact, some evidence that this is approximately the best possible prediction error for predicting MCAS scores. To examine this question, Feng, Heffernan & Koedinger (to appear) computed the split-half *Pct.Err* of the MCAS, using the MCAS scores of ASSISTments students, to be approximately 11%. Ayers & Junker (2006) derived a formula for the MSE (mean-square error) of prediction of one test from another, based on the classical true-score theory reliabilities of the two tests, and used this formula to bound the MAD. Using the published reliability of the 2005 MCAS exam, and the distribution of reliabilities of the various samples of ASSISTment questions seen by students, they estimated that the optimal MAD for predicting the 54-point raw MCAS score using classical true-score models should be no greater than approximately 5.21, or equivalently an optimal *Pct.Err* of about 9.65%. Thus the bound that we are reaching empirically in Table 3 may be close to the limit of what is possible, given the reliabilities of the MCAS and ASSISTment data.

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. In addition, the ASSISTment system has provided a very useful testbed for developing tutoring system architecture, authoring systems, and online cognitive modeling and prediction technologies.

This chapter has dealt primarily with this prediction function of the ASSISTments system. Our work has shown that a variety of prediction models can work well for this purpose. There is clearly a tradeoff between using cognitive/psychometric models that appropriately account for question difficulty, vs. using online prediction metrics measuring students’ efficiency and help-seeking behavior. The best approaches combine these two kinds of data.

Turning to teacher feedback, 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). Currently we are beginning to focus statistical modeling work on improving the modeling underlying these reports. For example, Cen, Koedinger & Junker (2006a) model learning curves using ideas of Draney, Pirolli and Wilson (1995) closely related to the logistic Rasch and LLTM models (equations (3) and (4) above). This approach can also be used to determine when the error rate on each KC is low enough that further practice is inefficient for the student (Cen, Koedinger & Junker, 2006b). Another approach combines the knowledge tracing algorithm of Corbett, Anderson & O’Brien (1995) with Bayes Net (DINA) models (Junker & Sijtsma, 2001); the key issue in deciding which approach to pursue will of course be model fit and interpretability.

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

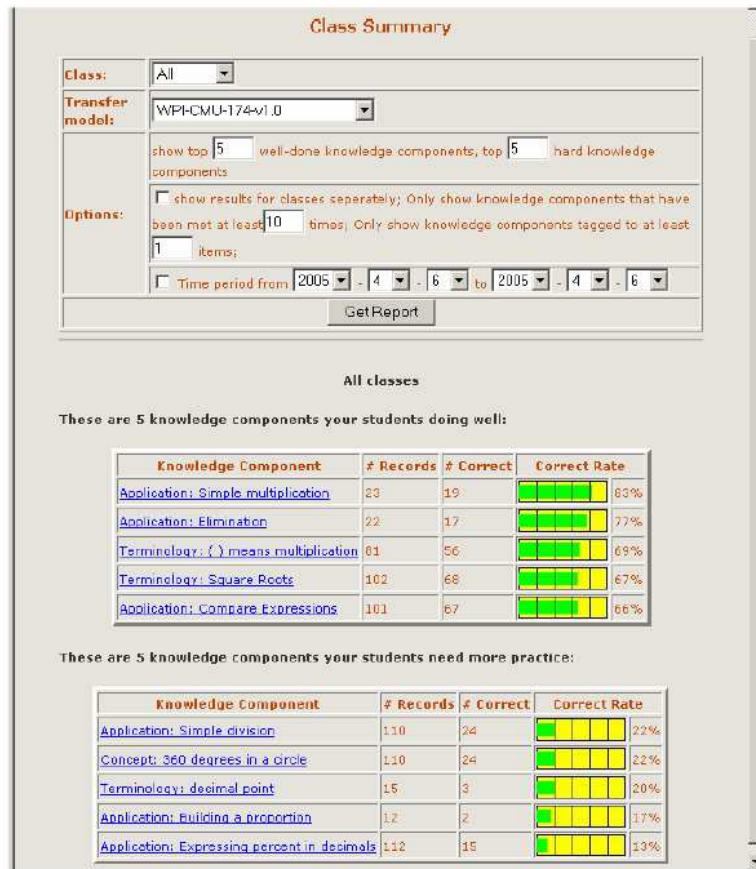


Figure 11: Classroom-level KC's report for teachers in the ASSISTment system. Student-level gradebook information is also available in the system.

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. As a way of managing this complexity, a mapping tool has been developed to help map KC's and groups of KC's in one transfer model to those in another transfer model. This is helpful, but it does not obviate the need to report to different stakeholders using different models of student proficiency.

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

- Aleven, V.A.W.M.M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2).
- Anozie, N. (2006). *Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system*. Proposal submitted to the National Council on Measurement in Education 2007 Annual Meeting.
- Anozie, N.O. & 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. & 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.
- Ayers, E. & Junker, B. W. (2006). *IRT modeling of tutor performance to predict end of year exam scores*. Working paper.
- Barnes, T. (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In the Proceedings of the AAAI-05 Workshop on Educational Data Mining, Pittsburgh, 2005 (AAAI Technical Report #WS-05-02).
- Campione, J.C., Brown, A.L., & Bryant, N.R. (1985). Individual differences in learning and memory. In R.J. Sternberg (Ed.). *Human abilities: An information-processing approach*, 103–126. New York: W.H. Freeman.

- Cen, H., Koedinger K., & Junker B. (2005). Automating Cognitive Model Improvement by A* Search and Logistic Regression. In *Technical Report (WS-05-02) of the AAAI-05 Workshop on Educational Data Mining, Pittsburgh, 2005*.
- Cen, H., K. Koedinger, and B. Junker (2006a). *Learning factors analysis: a general method for cognitive model evaluation and improvement*. Presented at the Eighth International Conference on Intelligent Tutoring Systems (ITS 2006), Jhongli, Taiwan.
- Cen, H., K. Koedinger, and B. Junker (2006b). *Is more practice necessary? Improving learning efficiency with the Cognitive Tutor through educational data mining*. Submitted to the 13th Annual Conference on Artificial Intelligence in Education (AIED 2007).
- 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, eds., *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.
- Corbett, A. T., Koedinger, K. R., & 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.
- Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. Nichols, S. Chipman, & R. Brennan, eds., *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.
- Embretson, S. E. (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. T., & Koedinger, K. R. (2006). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In Ikeda, Ashley & Chan (Eds.) *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp 31–40.
- Feng, M., Heffernan, N., Mani, M., & Heffernan, C. (2006). *Using mixed effects modeling to compare different grain-sized skill models*. AAAI06 Workshop on Educational Data Mining, Boston MA.
- Feng, M., Heffernan, N. T., & Koedinger, K. R. (in press). Addressing the testing challenge with a web-based E-assessment system that tutors as it assesses. *Proceedings of the 15th Annual World Wide Web Conference*. ACM Press (Anticipated): New York, 2005.

- Fischer, G.H. & Molenaar, I.W. (1995). *Rasch models: foundations, recent developments and applications*. New York: Springer-Verlag.
- Grigorenko, E. L. and Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124, 75–111.
- Junker, B.W. & Sijtsma K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. Invited paper in *Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Education)*.
- Koedinger, K. R., Anderson, J. R., Hadley, W. H., & 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., & Shapiro, L. J. (2000). *Carnegie Learning's Cognitive Tutor: Summary research results*. White Paper. Pittsburgh, PA: Carnegie Learning.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187–212.
- Mislevy, R. J. (1991). Randomization-based inferences about latent variables from complex samples. *Psychometrika*, 56, 177–190.
- Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In Laskey, K. B., and Prade, H., eds., *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, 437–446. S.F., Cal.: Morgan Kaufmann Publishers. See also <http://www.cse.ucla.edu/CRESST/Reports/TECH518.pdf>.
- Murphy, K. P. (2001). The Bayes Net Toolbox for MATLAB. *Computing Science and Statistics*, Vol. 33. Available online at <http://citeseer.ist.psu.edu/murphy01bayes.html>.
- Olson, L. (2005). Special report: testing takes off. *Education Week*, November 30, 2005, pp. 10–14. Also available on-line from <http://www.edweek.org/media/13testing.pdf>
- Pardos, Z. A., Feng, M., Heffernan, N. T. & Heffernan, C. L. (2006). *Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effect Methods*. Submitted to the the 13th Conference on Artificial Intelligence In Education (AIED 2007).
- Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. L. (2006). *Using Fine Grained Skill Models to Fit Student Performance with Bayesian Networks*. Workshop in Educational Data Mining held at the Eighth International Conference on Intelligent Tutoring Systems. Taiwan. 2006.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

- 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., Walonoski, J.A., Macasek, M.A., & Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th Artificial Intelligence In Education*. Amsterdam: ISO Press. pp 555–562.
- Razzaq, L., Feng, M., Heffernan, N. T., Koedinger, K. R., Junker, B., Nuzzo-Jones, G., Macasek, N., Rasmussen, K. P., Turner, T. E. & Walonoski, J. (to appear). A web-based authoring tool for intelligent tutors: blending assessment and instructional assistance. In Nedjah, N., et al. (Eds). *Intelligent Educational Machines*. Intelligent Systems Engineering Book Series (see <http://isebis.eng.uerj.br>).
- Rothman, S. (2001). *2001 MCAS Reporting Workshop: The second generation of MCAS results*. Massachusetts Department of Education. Downloaded November 2006 from http://www.doe.mass.edu/mcas/2001/news/reporting_wkshp.pps.
- Schofield, L., Taylor, L., & Junker, B. W. (2006). The use of cognitive test scores in evaluating black-white wage disparity. Working paper.
- Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Occurrence*. Oxford University Press, New York.
- Spiegelhalter, D. J., Thomas, A. & Best, N. G. (2003) *WinBUGS Version 1.4 User Manual*. Cambridge: Medical Research Council Biostatistics Unit.
- 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.
- The R Foundation. (2006). *The R Project for Statistical Computing*. Accessed November 2006 at <http://www.r-project.org/>: Author.
- van der Linden, Wim J. & Hambleton, Ronald K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Websites:

<http://www.assistment.org>
<http://www.learnlab.org>
<http://www.educationaldatamining.org>