

# Educational and Psychological Measurement

<http://epm.sagepub.com>

---

## **IRT Modeling of Tutor Performance to Predict End-of-Year Exam Scores**

Elizabeth Ayers and Brian Junker

*Educational and Psychological Measurement* 2008; 68; 972 originally published

online May 23, 2008;

DOI: 10.1177/0013164408318758

The online version of this article can be found at:

<http://epm.sagepub.com/cgi/content/abstract/68/6/972>

---

Published by:



<http://www.sagepublications.com>

**Additional services and information for *Educational and Psychological Measurement* can be found at:**

**Email Alerts:** <http://epm.sagepub.com/cgi/alerts>

**Subscriptions:** <http://epm.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

**Citations** <http://epm.sagepub.com/cgi/content/refs/68/6/972>

# IRT Modeling of Tutor Performance to Predict End-of-Year Exam Scores

Elizabeth Ayers

Brian Junker

*Carnegie Mellon University*

Interest in end-of-year accountability exams has increased dramatically since the passing of the No Child Left Behind Act in 2001. With this increased interest comes a desire to use student data collected throughout the year to estimate student proficiency and predict how well they will perform on end-of-year exams. This article uses student performance on the Assistment System, an online mathematics tutor, to show that replacing percentage correct with an Item Response Theory estimate of student proficiency leads to better fitting prediction models. In addition, it uses other tutor performance metrics to further increase prediction accuracy. Prediction error bounds are also calculated to attain an absolute measure to which the models can be compared.

**Keywords:** *cognitive modeling; Bayesian inference; intelligent tutoring systems; item response theory; reliability*

With the recent push in standardized testing in the United States, there has been an increased interest in predicting student performance on end-of-year exams from work done throughout the year (Olson, 2005). This has led to an increase in formative assessment and a growth of companies that provide assessment and prediction services (i.e., Pearson, <http://www.pearson.com>, and 4Sight, [www.cddre.org/Services/4Sight.cfm](http://www.cddre.org/Services/4Sight.cfm)). When predicting end-of-year exam performance, one of the most commonly used sources of student work is benchmark exams. Benchmark

**Authors' Note:** This Assistment Project was made possible by the U.S. Department of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant R305K03140, the Office of Naval Research grant N00014-03-1-0221, the Spencer Foundation, and NSF CAREER award to Neil Heffernan. Additional support for Ayers was provided by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 awarded to Carnegie Mellon University. This work would not have been possible without the assistance of the 2004-2005 WPI/CMU Assistment Team including Nathaniel O. Anozie, Andrea Knight, Ken Koedinger, Meghan Myers, Carolyn Rose all at CMU, Steven Ritter at Carnegie Learning, Mingyu Feng, Neil Heffernan, Tom Livak, Abraao Lourenco, Michael Macasek, Goss Nuzzo-Jones, Kai Rasmussen, Leena Razzaq, Terrence Turner, Ruta Upalekar, and Jason Walonoski all at WPI. Please address correspondence to Elizabeth Ayers, Department of Statistics, 132 Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213; e-mail: [eyayers@stat.cmu.edu](mailto:eyayers@stat.cmu.edu).

exams are typically paper-and-pencil exams given periodically throughout the year so teachers can get a snapshot of student knowledge at that time. A popular measure of student understanding for many researchers is percentage or number correct (e.g., Maccini & Hughes, 2000; Nuthall & Alton-Lee, 1995). Many popular prediction methods use a simple percentage correct or number of correct problems on the exams as a factor in prediction models (Bishop, 1998; Haist, Witzke, Quinlivan, Murphy-Spencer, & Wilson, 2003). This leads to linear prediction models of the form

$$Z_i = \lambda_0 + \lambda_1 \cdot \bar{X}_i + \sum_{m=2}^M \lambda_m \cdot Y_{im} + \varepsilon_i, \quad (1)$$

where  $Z_i$  is student  $i$ 's score on the end-of-year exam,  $\bar{X}_i$  is the percentage (or fraction) correct on the benchmark exam, and  $Y_{im}$  are other variables used in the regression such as subject or school-level background variables and other measures of performance. However, one drawback of this method is that it does not take into account the difficulty of the problems. For example, if two students see 10 different problems and both correctly answer 7, we should be cautious about using percentage (or number) correct to compare the students. If one set of problems is much harder than the other, then there is an obvious difference of abilities.

As a solution to this problem, one can use Item Response Theory (IRT; e.g., van der Linden & Hambleton, 1997), which relates student and problem characteristics to item responses. By separating the problem difficulty from student ability, we can estimate the student's true underlying ability no matter what set of problems are given. One of the simplest IRT models is the Rasch (Fischer & Molenaar, 1995), which models student  $i$ 's dichotomous response ( $0 = \text{wrong}$ ;  $1 = \text{correct}$ ) to problem  $j$ ,  $X_{ij}$ , in terms of student proficiency ( $\theta_i$ ) and problem difficulty ( $\beta_j$ ) as

$$P_j(\theta_i) = P(X_{ij} = 1 | \theta_i, \beta_j) = \frac{1}{1 + e^{-(\theta_i - \beta_j)}}. \quad (2)$$

When two students take different benchmark tests, the test characteristic functions (the average of the probabilities in equation 2,  $\bar{P}(\theta) = \frac{1}{J} \sum_{j=1}^J P_j(\theta_i)$ ) will be different, depending on the difficulty of the items in the two tests. Hence, the MLE  $\hat{\theta} = \bar{P}^{-1}(\bar{X})$  will automatically adjust estimated proficiency for the differing difficulty of the items on the two benchmark tests, even if  $\bar{X}$  is the same for both students. Thus, the IRT estimate of student proficiency is scaled according to the difficulty of the problems that the student saw.

One could then replace percentage correct in equation 1 with the estimated student proficiency to obtain

$$Z_i = \lambda_0 + \lambda_1 \cdot \theta_i + \sum_{m=2}^M \lambda_m \cdot Y_{im} + \varepsilon_i, \quad (3)$$

where  $Z_i$  and  $Y_{im}$  are the same as in equation 1 and  $\theta_i$  is student  $i$ 's estimated IRT proficiency. This approach is similar to the IRT-based errors-in-variables regression model used by Schofield (2007) in public policy.

In this article, we illustrate the steps described above using data from an online mathematics tutor known as the Assistment System (Heffernan, Koedinger, & Junker, 2001; Junker, in press). During the 2004-2005 school year, more than 900 eighth-grade students in Massachusetts used the tutor to prepare for the Massachusetts Comprehensive Assessment System (MCAS) exam. The MCAS exam is part of the accountability system that Massachusetts uses to evaluate schools and satisfy the requirements of the 2001 No Child Left Behind Act (see more at <http://www.doe.mass.edu/mcas>). In this analysis, the benchmark exams are the unique set of tutor problems that each student received and the other variables,  $Y_{im}$ , in equation 3 are other manifest measures of student performance such as number of hints asked for and time spent answering problems.

We also compare the prediction models we construct with one another. To compare models, we computed the 10-fold cross-validation mean absolute prediction error, or the mean absolute deviation (MAD), shown in equation 4. MAD is used because it is considered to be more interpretable by the Assistment developers. We also report the cross-validation mean square error (MSE).

$$\text{MAD} = \text{mean}|Z_i - \text{predicted } Z_i| = \frac{1}{N} \sum_{i=1}^N |Z_i - \text{predicted } Z_i|. \quad (4)$$

In equation 3, there are many different variables  $Y_{im}$  that can be used and many different choices of IRT models to estimate student proficiency. By comparing the prediction error of these models, we can tell when one model is doing better than another, but we cannot tell whether any one model is doing well or poorly in an absolute sense. We will use classical test theory (Lord & Novick, 1968) to obtain approximate best-case bounds on the prediction error in terms of the reliabilities of the individual benchmark tests taken by the students. This gives us an absolute criterion against which to compare the prediction error of various models. If the prediction error of a model is larger than the upper bound, we know to throw out the model and search for a better one.

Each year, the 200 to 280 reporting scale used to communicate MCAS results to the public is recalculated by first using a standard-setting procedure to set the achievement levels in terms of the raw number correct and then using piecewise linear transformations to turn the number-correct scores into values within the 200-to-280 range. This second step is done so that the reporting scale achievement level cut points remain the same from year to year (Rothman, 2001). We predict the raw number-correct score, 0 to 54, to avoid the artificial year-to-year variation introduced by this standard-setting and transformation process.

The study and data that this article uses are described in the following section. We then describe the statistical methods used to model student proficiency and

summarize the results. Classical test theory is then used to discuss how well we expect to do with predictions. Next, we look at several MCAS exam score prediction models and compare results. Finally, we offer some overall conclusions.

## The Assistment Project

### Design

During the 2004-2005 school year, more than 900 eighth-grade students in Massachusetts used the Assistment System. Eight teachers from two middle schools participated, with students using the system for 20 to 40 minutes every 2 weeks. Almost 400 main questions were randomly given to students in the Assistment System. The pool of main questions was restricted in various ways, for example, by the rate at which questions in different topic areas were developed for the tutor by the Assistment Project team and by teachers' needs to restrict the pool to topics aligned with current instruction. Thus, coverage of topics was not uniform, and students might see the same Assistment tasks more than once.

### Data

Students using the Assistment System are presented with problems that are either previously released MCAS exam questions or that are *prima facie* equivalent "morphs" of released MCAS exam questions; these are called "main questions." In other contexts (e.g., Embretson, 1999), item morphs are called "item clones." If students correctly answer a main question, they move on to another main question. If students incorrectly answer the main question, they are required to complete scaffolding questions that break the problem down into simpler steps. Students may make only one attempt on the main question each time it is presented but may take as many attempts as needed for each of the scaffolds. Students may also ask for hints if they get stuck answering a question.

The analysis in this article includes only those students who have MCAS exam scores recorded in the database. This narrows the sample size to 683 students. There are 354 different main questions seen by these students. Individual students saw between 1 and 252 problems; however, the distribution is right skewed, with a median of 71 problems and first and third quartiles of 39 and 107 problems, respectively. Previously, Farooque and Junker (2005) found evidence that skills behave differently in Assistment main questions and scaffolds. Because we want to make comparisons to the MCAS exam, the only Assistment data used in the IRT models are of performance (correct or incorrect) on Assistment main questions.

## IRT Model Estimation

Because performance on any particular problem depends on both student proficiency and problem difficulty, we use IRT models to factor out student proficiency and directly model problem difficulty. MCAS multiple-choice questions are scaled (Massachusetts Department of Education, 2004) for operational use with the 3-Parameter Logistic (3PL) model, and short-answer questions are scaled using the 2-Parameter Logistic (2PL) model from IRT (van der Linden & Hambleton, 1997). We know that Assistent main questions are built to parallel MCAS exam questions, so it might be reasonable to model Assistent main questions using the same IRT models. However, for simplicity the Rasch model, equation 2, was used. There is evidence that student proficiencies and problem difficulties have similar estimates under the 3PL and the Rasch model (Wright, 1995), so we are not losing much information by starting with the Rasch model. Note that in the Rasch model, the problem difficulty parameters  $\beta_j$  are not constrained in any way.

In our analysis, we consider  $N = 683$  students' dichotomous answers to up to  $J = 354$  Assistent main questions. There are many missing values because no student saw all of the problems. We treat these missing values as missing completely at random (MCAR), because problems were assigned to students randomly by the Assistent software from a "curriculum" of possible questions designed for all students by their teachers in collaboration with project investigators.

The dichotomous responses  $X_{ij}$  are modeled as Bernoulli trials:

$$X_{ij} \sim \text{Bern}(P_j(\theta_i)),$$

where  $i = 1, \dots, N$ ;  $j = 1, \dots, J$ ; and  $P_j(\theta_i)$  is given as above by equation 2. Under the usual IRT assumption of independence between students and between responses, given the model parameters, the complete data likelihood can be written as

$$P(X=x) = \prod_{i=1}^N \prod_{j:i \text{ saw } j} P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}. \quad (5)$$

We estimated the student proficiency ( $\theta_i$ ) and problem difficulty ( $\beta_j$ ) parameters using Markov Chain Monte Carlo methods with the program WinBUGS (Bayesian inference Using Gibbs Sampling; Spiegelhalter, Thomas, & Best, 2003; WinBUGS and R code available from the authors on request). The Rasch model, equations 2 and 5, was estimated using the data with the priors  $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$  and  $\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$ . We placed a weak normal hyperprior on  $\mu_\beta$  and a weak inverse-Gamma hyperprior on  $\sigma_\beta^2$ . In item response models, the location and scale of the latent variable, and hence of problem difficulty parameters, are not fully identified, which can undermine comparisons between fits on different data sets. We decided to fix the (prior) mean and variance of the student proficiency ( $\theta$ ) to be 0.69 and 0.758. These values were found by preliminary analysis using weak hyperpriors on these parameters. All estimates mentioned herein refer to the posterior means of the parameters.

Before moving on, it is worth mentioning three reasons why we chose to implement MCMC estimation over maximum likelihood methods. First, in a research setting where we are combining IRT and regression methods, as in equation 3, we are willing to trade speed of joint maximum likelihood (JML) or marginal maximum likelihood (MML) estimation for ease of implementation of MCMC. To set up the MCMC estimation, we only need to specify the likelihood and prior distributions for the parameters, versus calculating the first and second derivations of the likelihood. Second, we can make more complete uncertainty calculations within the Bayesian framework. Finally, depending on how the MCMC output is utilized, the asymptotic properties of the MCMC estimates will behave like either JML or MML estimates (Patz & Junker, 1999). As a comparison, we calculated the MML estimates using ConQuest (Wu, Adams, & Wilson, 1999). For the majority of the problems, the 95% confidence intervals for the MCMC and MML estimates overlapped quite well. However, for 4.5% of the problems the ConQuest estimates were unreasonably extreme (indicating lack of MML convergence) compared with the MCMC estimates.

To explore the fit of the Rasch model, we looked at the per-problem standardized residuals:

$$r_j = \frac{n_j - E(n_j)}{\sqrt{\text{Var}(n_j)}}, \tag{6}$$

where  $n_j = \sum_{i:i \text{ saw } j} X_{ij}$  is the number of correct answers to problem  $j$ ,  $E(n_j)$  is its expected value estimated from fitting the model in equation 2, and  $\text{Var}(n_j)$  is its variance estimated from the same model. Because these residuals are standardized, we expect the majority to fall between  $-3$  and  $3$ . The plot on the left in Figure 1 shows these standardized residuals. One can note that they fall between  $-0.6$  and  $1.4$ , indicating a good fit.

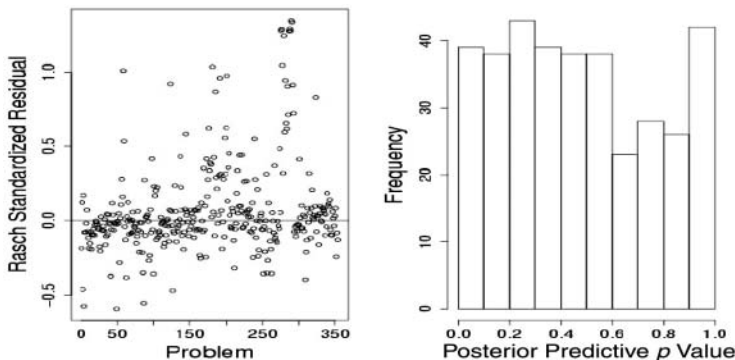
We also calculated the per-problem outfit statistics (van der Linden & Hambleton, 1997, p. 113),

$$T_j(x|\phi) = \sum_{i=1}^{N_j} \frac{(x_{ij} - E_{ij})^2}{N_j W_{ij}},$$

where  $N_j$  is the number of students that saw problem  $j$ ,  $x_{ij}$  is student  $i$ 's response on problem  $j$ ,  $E_{ij}$  is the expected value of  $X_{ij}$  conditional on the parameter vector  $\phi$ , and  $W_{ij}$  is the variance of  $X_{ij}$  also conditional on  $\phi$ . To check the per-problem fit of each model, the posterior predictive  $p$  (PPP) value (Gelman, Carlin, Stern, & Rubin, 2004)—the expected value of the classical  $p$  value over the posterior distribution of the parameter vector given the model and the observed data—was estimated using

$$p_i \approx \frac{\#\{s : T_i(x|\phi_x) < T_i(x^*|\phi_x); s = 1, 2, \dots, M\}}{M},$$

**Figure 1**  
**Standardized Residuals and Posterior Predictive  $p$  Values**



which compares the observed values of the test statistic to values of the test statistic for data simulated from the model. For this calculation, the simulated data ( $x^*$ ) were obtained by using the Markov Chain given by WinBUGS. Similar to classical  $p$  values, there is reason to question the fit of the model to problem  $i$  if  $P_i$  is small. A weakness of the PPP value is that it uses the data twice, once to calculate the observed test statistics and again to simulate data to calculate the PPP value. One consequence of this is that PPP values are not uniformly distributed and tend to be conservative (Gelman, Meng, & Stern, 1996, p. 790). However, we can still expect the PPP values to aggregate around zero if there is serious misfit for some of the problems. We can see that the histogram of PPP values on the right in Figure 1 is roughly uniform, which we would expect if the model fit is acceptable.

We also considered the Linear Logistic Test Model (LLTM; Fischer, 1974) and random-effects LLTM (Janssen & De Boeck, 2006), but the fit of both models was inadequate in comparison with the Rasch model. More information can be found in previous work (Ayers & Junker, 2006).

### Reliability and Predictive Accuracy

Before exploring the predictive accuracy of our models using the MAD measure defined in equation 4, it is important to ask how well Assistent scores could predict MCAS scores under ideal circumstances. Let us begin by assuming the MCAS

exam and the Assisment System to be two parallel tests of the same underlying construct. Following classical test theory (Lord & Novick, 1968), we have

$$\begin{aligned} X_{i1} &= T_i + \varepsilon_{i1}, \\ X_{i2} &= T_i + \varepsilon_{i2}, \end{aligned}$$

where the true score of student  $i$  is  $T_i$ ,  $X_{it}$  is student  $i$ 's observed score on test  $t$ , and  $\varepsilon_{it}$  is the error on test  $t$ .

We have followed the usual assumptions that the expected value of the error terms are zero, the error terms are uncorrelated, and that the error terms and the true score are uncorrelated. The expected mean square error (MSE) between the tests is then

$$E[(X_{i1} - X_{i2})^2] = E[(\varepsilon_{i1} - \varepsilon_{i2})^2] = \sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2.$$

Because the reliability of test  $t$  ( $t = 1$  or  $2$ ) is defined as

$$r_t = \frac{\sigma_T^2}{\sigma_{X_t}^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{\varepsilon_t}^2}, \tag{7}$$

some algebra then shows that the root mean square error (RMSE) is

$$\text{RMSE} = \sqrt{E[(X_{i1} - X_{i2})^2]} = \sigma_T \sqrt{\left(\frac{r_1 + r_2}{r_1 \cdot r_2} - 2\right)}.$$

This can be converted into lower and upper bounds on the MAD score as follows. Using the Cauchy-Schwarz inequality for Euclidean spaces (Protter & Morrey, 1991, p. 130) with  $x_i = |\text{MCAS}_i - \text{predicted MCAS}_i|$  and  $Y_i = 1$ ,

$$\sum_{i=1}^N |\text{MCAS}_i - \text{predicted MCAS}_i| \leq \sqrt{N} \cdot \sqrt{\sum_{i=1}^N (\text{MCAS}_i - \text{predicted MCAS}_i)^2}.$$

We can then scale both sides by  $1/N$  to achieve

$$\text{MAD} \leq \text{RMSE}.$$

We can also bound the MAD from below. First, let  $x_i = \text{MCAS}_i - \text{predicted MCAS}_i$  and  $|x_{\max}|$  denote the absolute maximum deviation between the true and predicted MCAS scores. Then,

$$\text{RMSE}^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 \leq \frac{1}{N} \sum_{i=1}^N |x_i| \cdot |x_{\max}| = |x_{\max}| \frac{1}{N} \sum_{i=1}^N |x_i| = |x_{\max}| \text{MAD},$$

so we have that

$$\frac{1}{|x_{\max}|} \cdot \text{RMSE}^2 \leq \text{MAD}.$$

Thus, our lower and upper bounds for the MAD score are

$$\frac{1}{|x_{\max}|} \cdot \text{RMSE}^2 \leq \text{MAD} \leq \text{RMSE}. \quad (8)$$

From equation 7, we have that  $\sigma_T^2 = r_T \cdot \sigma_X^2$ . In the most recent technical report published (Massachusetts Department of Education, 2004), the MCAS has listed  $r_{i=1} = 0.9190$  and  $\sigma_X^2 = 142.39$ , so that in predicting MCAS exam scores from Assistent scores we have

$$\text{RMSE} = \sqrt{130.86 \times \left( \frac{0.9190 + r_2}{0.9190 \cdot r_2} - 2 \right)}, \quad (9)$$

where  $r_2$  is the reliability of the Assistent score.

However, because each student completes a unique set of Assistent main questions, we could not calculate  $r_2$  directly. Instead, we calculated reliability separately for each student. For this purpose we considered a reduced data set of 616 students who had 10 or more problems completed for which all pairs of correlations were available. To estimate the per-student reliability, we used Cronbach's alpha coefficient (Cronbach, 1951),

$$\alpha_i = \frac{n_i \bar{r}_i}{1 + (n_i - 1) \bar{r}_i}. \quad (10)$$

In equation 10,  $n_i$  is the number of problems that student  $i$  saw and  $\bar{r}_i$  is the average interitem correlation for problems seen by student  $i$ . Once the per-student reliabilities were calculated, the per-student estimated RMSE values were computed using equation 9. Figure 2 shows the estimated reliabilities for the students who met the criteria explained above. It is interesting to note that the estimated RMSE is never lower than 4.44.

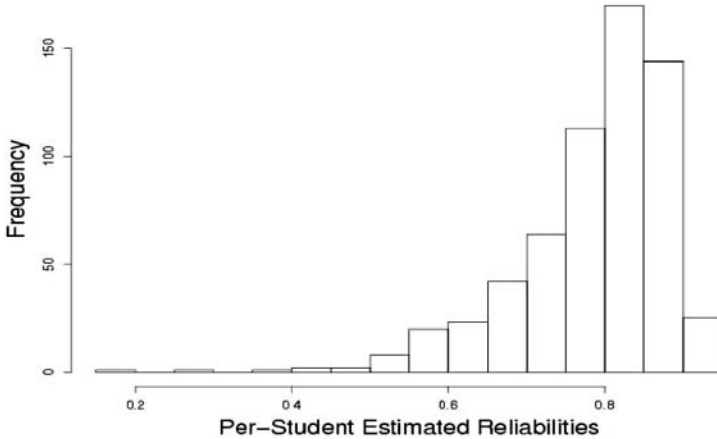
To have a single set of approximate bounds for the MAD score in equation 8, we found the median Assistent reliability, 0.8080, and the corresponding RMSE of 6.529 from equation 9. The largest deviation,  $|x_{\max}|$ , between the true and predicted MCAS scores among the models in Table 1 was 40.5. Substituting these values for RMSE and  $|x_{\max}|$  in equation 8, we find the approximate bounds,

$$1.053 \leq \text{MAD} \leq 6.529.$$

## MCAS Exam Score Prediction

We now combine student proficiencies estimated from a successful IRT model with other Assistent performance metrics to produce an effective prediction function,

**Figure 2**  
**Histogram of Per-Student Assistent Reliabilities Given by Equation 10**



following the work of Anozie and Junker (2006), using an errors-in-variables regression approach similar to that of Schofield (2007). The linear model is

$$MCAS_i = \lambda_0 + \lambda_1 \cdot \theta_i + \sum_{m=2}^M \lambda_m \cdot Y_{im} + \varepsilon_i,$$

where  $\theta_i$  is the proficiency of student  $i$  as estimated by the IRT model and  $Y_{im}$  is performance of student  $i$  on manifest measure  $m$ . WinBUGS was again used to find Bayesian estimates of the linear regression coefficients.

In practice, Assistent items will be calibrated and only  $\theta$  will need to be estimated. Thus, when estimating each of the following models, the IRT item parameters were fixed at their estimates from before, but student proficiency was reestimated. Because the measurement error for each person is small (ranging from 0.18 to 0.86 for the 683 students), it is tempting to plug in the  $\hat{\theta}$ s from before as well. However, a simulation study by Zwinderman (1991) showed an increased bias when  $\theta$  is replaced by  $\hat{\theta}$  as a dependent variable, and we expect the same results whenever  $\theta$  is used as an independent variable.

To compare the prediction models, we calculated the 10-fold cross-validation (CV) MAD score (equation 4). In K-fold CV, the data set is randomly divided into K subsets of approximately equal size. One subset is omitted (referred to as the testing set) and the remaining K-1 subsets (referred to as the training set) are used to fit the model. The fitted model is then used to predict the MCAS exam scores for the testing set. The desired statistic, in this case the MAD score, is then calculated

**Table 1**  
**Prediction Models**

Model	Variables	No. of Variables	CV MAD	CV RMSE	Notes
Model 1	Percentage correct on main questions	1	7.18	8.65	
Model 2	Rasch student proficiency	1	5.90	7.18	
Model 3 (Anozie & Junker, 2006)	Percentage correct on main questions and four other manifest performance measures	35	5.46	7.00	Uses multiple monthly summaries
Model 4	Rasch student proficiency and same four manifest performance measures as Model 3	5	5.39	6.56	Uses only year-end aggregates
Model 5	Rasch student proficiency and five manifest performance measures (one overlap with Models 3 and 4)	6	5.24	6.46	Optimized for student proficiency

Note: CV = cross validation; MAD = mean absolute deviation; RMSE = root mean square error.

for the testing set. The process is repeated  $K$  times (the folds), with each of the  $K$  subsets being used exactly once as the testing set. The  $K$  results from the folds are then averaged to produce a single estimation of the MAD score. By using cross-validation, we avoid using the data to both fit the model and give an estimate of the fit. In addition, we also report the 10-fold cross-validation root sample mean square error of the models.

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^N (MCAS_i - \text{predicted } MCAS_i)^2}$$

Table 1 shows results from several prediction models. Column 2 lists which variables are in the model. (For a full list and description of the variables used in each model, see Table 2.) Column 3 simply states the number of variables in the

**Table 2**  
**Definitions of Variables Used in Prediction Models**

Variable Name	Model	Definition
Student proficiency	2, 4, 5	IRT estimate of student proficiency
PctCorMain	1, 3	Percentage of correctly answered main questions
PctCorScaf	3, 4	Percentage of correctly answered scaffolds
SecIncScaf	3, 4	Number of seconds spent answering all incorrect scaffolds
NumPmAllScaf	3, 4, 5	Number of scaffolds completed per minute
NumHintsIncMainPerMain	3, 4	(Number of hints + number of incorrect main questions)/Number of main questions attempted
SecCorScaff	5	Number of seconds spent answering all correct scaffolds
SecIncMain	5	Number of seconds spent on incorrect main questions
MedSecIncMain	5	Median number of seconds per incorrect main question
PctSecIncMain	5	Percentage of time on main questions spent on incorrect main questions

Note: IRT = Item Response Theory.

model. Columns 4 and 5 give the CV MAD score and the CV RMSE respectively. Column 6 offers some important notes about the model. Historically, and in particular within the Assistentment Project, percentage correct on main questions has been used as a proxy for student ability. To see if any information was gained by simply using the Rasch estimate of student proficiency, we compared the two models with only these variables. Model 1 is the simple linear regression using only percentage correct on main questions and has a MAD score of 7.18. Model 2 uses only the Rasch student proficiency and gives a MAD score of 5.90. By simply using IRT to account for problem difficulty in estimating student proficiency, we can drop the MAD score a full point. Accounting for problem difficulty gives a more efficient estimate of how well a student is doing and leads to better predictions. Model 3, from Anozie and Junker (2006), uses as predictors monthly summaries from October to April for percentage correct on main questions and four other manifest measures of student performance. Model 4 uses the year-end aggregates of the same variables and substitutes Rasch student proficiency for percentage correct on main questions. We see that Model 4 gives a slightly lower MAD score. Thus, by using Rasch student proficiency (in place of percentage correct on main questions) we can use fewer measures of student performance on Assistentment problems.

Model 5 was optimized (for MAD score) for Rasch student proficiency and year-end aggregates of student performance measures using backwards variable selection implemented in WinBUGS and R (R Development Core Team, 2004; WinBUGS and R code available from the authors on request). To start, we used the same 12 variables as Anozie and Junker (2006), excluding percentage correct on

main questions and adding Rasch student proficiency. We ran the full model and all models excluding one variable, with the caveat that student proficiency was always kept in the model. For each model, MCAS exam scores were predicted and MAD scores calculated. The model with the lowest MAD score was then used as the new “full” model. This process was repeated until removing variables from the full model no longer reduced the MAD score. The final model, which contained student proficiency and five manifest measures of student performance, gave a MAD score of 5.24, a slight improvement from Model 4. Overall, the ability to use fewer variables makes the effort expended in estimating the IRT models worth it.

The regression equation for Model 2 is

$$\text{MCAS}_i = 18.289 + 10.425 \cdot (\text{Rasch Student Proficiency}). \quad (11)$$

From this, we see that there is a baseline MCAS exam score prediction of 18 points and for each additional unit of estimated Rasch student proficiency, we add 10.425 to the exam score prediction. As a student’s proficiency increases, so does his or her exam score prediction. The regression equation for Model 5 is

$$\begin{aligned} \text{MCAS}_i = & 8.514 + 10.336 \cdot (\text{Rasch Student Proficiency}) \\ & + 8.928 \cdot (\text{NumPmAllScaf}) + 0.004 \cdot (\text{SecCorScaff}) \\ & + 0.032 \cdot (\text{MedSecIncMain}) - 0.001 \cdot (\text{SecIncMain}) \\ & - 2.696 \cdot (\text{PctSecIncMain}). \end{aligned} \quad (12)$$

In equation 12, the increase in MCAS score for each unit of increase in Rasch proficiency is about the same as in equation 11. However, the baseline of 18.289 has been decomposed into a new baseline of about 8.5 points, incremented or decremented according to various measurements of response efficiency. The largest increment, 8.928, comes from the rate at which scaffolding questions are completed and the largest decrement, 2.696, comes from time spent on answering main questions incorrectly.

Now that we have compared models to one another, we need to compare the models to the bounds calculated above. Recall from the previous section that we have a bound of

$$1.053 \leq \text{MAD} \leq 6.529.$$

From Table 1, one can see that Model 5 has a MAD score of 5.24, which is well below the upper bound.

Moreover, the RMSE reported for Model 5, 6.46, is similar to our estimated optimal RMSE of 6.53. It should also be noted that with a perfect Assistent reliability in equation 9, the estimated RMSE would be 5.576 and the bound would be

$$0.768 \leq \text{MAD} \leq 5.576.$$

Again, the MAD score of Model 5 is below this upper bound. Using a split-half reliability calculation on the MCAS exam itself, Feng, Heffernan, and Koedinger (2006) found an average MAD score of 5.94. Because we are achieving MAD scores less than this and the two previously mentioned upper bounds, we do not expect to do much better without an increase in the reliability of the MCAS exam.

## Discussion

In this article, we have developed a framework to create prediction functions for end-of-year exam scores using an IRT estimate of student ability based on work done throughout the school year. Although this framework was illustrated using data from an online mathematics tutor, other benchmark work, such as homework or paper-and-pencil exams, could be used to predict end-of-year exam scores as well.

In addition to developing this general framework, our research generated an additional finding. Prediction using IRT scores is more effective than prediction using number-correct scores. For example, the predictions based on our Rasch model always produced lower MAD and RMSE prediction errors than the corresponding predictions based on number-correct scores. Moreover, the IRT-based predictions were essentially as good as one could do with parallel tests, even though our Assistent System was not constructed to be parallel (in the classical test theory sense) to the MCAS exam.

## References

- Anozie, N. O., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In *Proceedings of the American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA* (Technical Report WS-06-05, pp. 1-6). Menlo Park, CA: AAAI Press.
- Ayers, E., & Junker, B. W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In *Proceedings of the American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA* (Technical Report WS-06-05, pp. 14-20). Menlo Park, CA: AAAI Press.
- Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *The Journal of Economic Education, 29*(2), 171-182.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.
- Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*, 407-433.
- Farooque, P., & Junker, B. W. (2005). *Behavior of skills within MCAS and Assistent main problems*. Final project poster, Department of Statistics, Carnegie Mellon University.
- 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 M. Ikeda, K. Ashley, & T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 31-40). Berlin: Springer-Verlag.

- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen* [Introduction to the theory of psychological tests: Foundations and applications]. Bern, Switzerland: Verlag Hans Huber.
- Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies [with discussion]. *Statistica Sinica*, 6, 733-807.
- Haist, S. A., Witzke, D. B., Quinlivan, S., Murphy-Spencer, A., & Wilson, J. F. (2003). Clinical skills as demonstrated by a comprehensive clinical performance examination: Who performs better—men or women? *Advances in Health Sciences Education*, 8, 189-199.
- Heffernan, N. T., Koedinger, K. R., & Junker, B. W. (2001). *Using Web-based cognitive assessment systems for predicting student performance on state exams*. Research proposal to the Institute of Educational Statistics, U.S. Department of Education. Department of Computer Science at Worcester Polytechnic Institute, Worcester County, MA.
- Janssen, R., & De Boeck, P. (2006). *A random-effects version of the LLTM*. Technical report, Department of Psychology, University of Leuven, Belgium.
- Junker, B. W. (in press). Using on-line tutoring records to predict end-of-year exam scores: Experience with the ASSISTments Project and MCAS 8th grade mathematics. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting*. Maple Grove, MN: JAM Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Maccini, P., & Hughes, C. A. (2000). Effects of a problem-solving strategy on the introductory algebra performance of secondary students with learning disabilities. *Learning Disabilities Research and Practice*, 15, 10-21.
- Massachusetts Department of Education. (2004). *2004 MCAS technical report*. Retrieved December 2005 from <http://www.doe.mass.edu/mcas/2005/news/04techrpt.pdf>
- Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal*, 31(1), 185-223.
- Olson, L. (2005, November 30). State test programs mushroom as NCLB mandate kicks in. *Education Week*, pp. 10-14.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146-178.
- Protter, M. H., & Morrey, C. B., Jr. (1991). *A first course in real analysis* (2nd ed.). New York: Springer.
- R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing (ISBN 3-900051-07-0). Available from <http://www.R-project.org>
- Rothman, S. (2001). *2001 MCAS reporting workshop: The second generation of MCAS results*. Massachusetts Department of Education. Retrieved January 2007 from [http://www.doe.mass.edu/mcas/2001/news/reproting\\_wkshp.pps](http://www.doe.mass.edu/mcas/2001/news/reproting_wkshp.pps)
- Schofield, L. (2007, June 10). *Using cognitive test scores in social science research*. Thesis proposal, Department of Statistics, Carnegie Mellon University.
- Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2003). *WinBUGS: Bayesian inference using Gibbs sampling, manual version 1.4*. Cambridge, UK: Medical Research Council Biostatistics Unit.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

- Wright, B. D. (1995). 3PL or Rasch? *Rasch Measurement Transactions*, 9(1), 408-409.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalised item response modelling software, manual draft release 2*. Hawthorn, Australia: Australian Council for Educational Research.
- Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, 56, 589-600.