Do skills combine additively to predict task difficulty in eighth-grade mathematics?

Elizabeth Ayers, Brian Junker
Department of Statistics, Carnegie Mellon University
Pittsburgh, PA, USA
{eayers, brian} @ stat.cmu.edu

Abstract

During the 2004–2005 school year, over 900 eighth-grade students used an online intelligent tutoring system, the Assistment System of Heffernan, et al. (2001), to prepare for the mathematics portion of the Massachusetts Comprehensive Assessment System (MCAS) end-of-year exam. A transfer model, identifying the skills that each tutoring task and exam problem depends upon, was developed to help align tutoring tasks with exam problems. We use a Bayesian form of item response theory (IRT) modeling to attempt to model the difficulty of tutoring tasks and exam items additively in terms of these component skills: the more skills, the more difficult the task or test item. Our goal is to directly examine the alignment between tutoring tasks and assessment items and to use the transfer model to build more efficient functions for predicting end-of-year exam performance from student activity with the online tutor. However, our analysis shows that the additive skills model (the Linear Logistic Test Model, LLTM) does not adequately account for task-to-task or item-to-item variation in difficulty.

Keywords: Cognitive modeling, Bayesian inference, intelligent tutoring systems, item response theory.

1 Introduction

The Assistment[1] Project is a collaboration between the Computer Science Department at Worcester Polytechnic Institute and several departments at Carnegie Mellon University. The overall goal of the project is to build a reliable on-line tutor, referred to as the Assistment System, to prepare students for the Mathematics portion of 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 NCLB law[2]. In addition, we would like to be able to predict students' performance on the MCAS exam from their performance on the Assistment System and provide reliable feedback to teachers about student knowledge.

Recently much work has been done by our colleagues in the Assistments Project to predict MCAS scores, e.g. from monthly aggregates of dynamic tutoring metrics (Anozie & Junker, 2006), from a detailed Bayes net specification of student skills (Pardos, Heffernan, Anderson & Heffernan, 2006), or from linear growth curve models for student performance (Feng, Heffernan & Koedinger, 2006; Feng, Heffernan, Mani & Heffernan, 2006). Although this work is very promising, it has been difficult to reduce the mean absolute prediction error below about 10% of the total possible MCAS score, which is still somewhat high for reliable and accurate prediction.

[1] The term "Assistment" was coined by Kenneth Koedinger and blends Assessment and Assisting.

[2] See more on http://www.doe.mass.edu/mcas.

One possible impediment to further reducing prediction error is a lack of alignment between the way skills contribute to task difficulty in the Assistment tutoring system, vs. how they contribute to item difficulty in the MCAS exam. The work in this paper uses Item Response Theory (IRT; e.g. van der Linden & Hambleton, 1997) to model task and test item difficulty additively in the skills required for the items: the more skills required for each task or test item, the more difficult the item is expected to be. The additive model we use, the Linear Logistic Test Model (LLTM; Fischer, 1974; van der Linden & Hambleton, Chapter 13), effectively constrains task difficulties according to a transfer model. This constrained model is compared with a Rasch IRT model to see if the constraints implemented in the transfer model adequately model task-to-task or item-to-item variation in difficulty (this is conceptually similar to computing R^2 as the "proportion of variance explained" in linear regression but there is no simple analogue to R^2 in our setting). Additive-difficulty models have been used successfully to model intelligent tutoring data in other settings (eg. Draney, Pirolli & Wilson, 1995); such models should be distinguished, e.g., from conjunctive Bayes net models (e.g. Pardos et al., 2006; or Junker & Sijtsma, 2001) which focus on student performance rather than task difficulty.

The study and data on which this paper is based are described in Section 2. Section 3 gives more insight into why we should assess the coding of skills to problems. In Section 4 we describe the statistical methods used and summarize our results. We then present a random effects model in Section 5 in an attempt to find a better fitting model. Finally, we compare skills in Assistment main questions and MCAS items in Section 6, and offer some conclusions in Section 7.

2 The Study

2.1 Design

During the 2004–2005 school year, over 900 8th grade students in Massachusetts used the Assistment System. Eight teachers from two different middle schools participated, with students using the System for 20–40 minutes every two weeks. There were almost 400 main questions in the Assistment System which were randomly given to students. 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 Assistments 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.

2.2 Data

Students using the Assistment System are presented with problems that are either previously released MCAS exam items or that are *prima facie* equivalent "morphs" of released MCAS exam items; these are called

"main questions". If students correctly answer a main question, they move onto another main question. If students incorrectly answer the main question, they are required to complete scaffolding questions which break the problem down into simpler steps. Students may make only one attempt on the main question each time that 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 in answering a question.

The analysis in this paper includes only those students which have MCAS exam scores recorded in the database. This narrows the sample size to a total of 683 students. Previously Farooque & Junker (2005) found evidence that skills behave differently in Assistment main questions and scaffolds. Since we want to make comparisons to the MCAS exam, the only Assistment data that is utilized in this paper is performance (correct/incorrect) on Assistment main questions. There are a total of 354 different main questions seen by the above students. Also analyzed in this paper is performance on the 39 Spring 2005 MCAS exam questions for the 683 students whose Assistment performance was used.

To model and predict difficulty of items, we need a measure of what skills items do and do not contain. We can break the problems down into individual mathematics skills and record the dependencies between problems and skills. The skill indications can be assembled into a transfer model. The transfer model, also referred to as a Q-matrix (Embretson, 1984; cf. Barnes 2002 for a recent, more-elaborate application in intelligent tutoring) or skill coding, is a matrix

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ \vdots & \ddots & & \vdots \\ q_{J,1} & q_{J,2} & \cdots & q_{J,K} \end{bmatrix},$$

where $q_{j,k} = 1$ if problem j contains skill k and 0 if it does not. Thus, the transfer model simply indicates which skills each problem contains.

Currently attention is focused on a single transfer model known as the WPI-April-2005 transfer model. In the remainder of this paper, unless otherwise noted, any reference to a transfer model refers to this particular model. The current version of this model contains a total of 106 skills, 77 of which appear on the Assistment main questions included in this analysis and 40 of which appear on the Spring 2005 MCAS exam. Neither of these skill sets is a subset of the other.

3 Research Question - Transfer Model Assessment

Our goal is to use the transfer model to directly examine the alignment between tutoring tasks and assessment items, with an eye toward building more efficient functions for predicting end-of-year exam performance from student activity with the online tutor. If skills function differently for Assistment main questions and MCAS test items, then additional adjustments will need to be made to improve functions predicting MCAS performance from student activity with the tutor. Since performance on any particular test item depends on

both student proficiency and item difficulty, we use a member of the family of IRT models, the LLTM, to factor out student proficiency and directly model item difficulty as a function of skills required for the item.

4 Assessment of the WPI-April-2005 Transfer Model

MCAS multiple choice questions are scaled[3] using 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 Assistment Items are built to parallel MCAS items and so it would be reasonable to model Assistment Items using the same IRT models. However, for simplicity, the Rasch model,

$$P_j(\theta_i) = P(X_{i,j} = 1|\beta_j, \theta_i) = \frac{1}{1 + e^{-(\theta_i - \beta_j)}}, \tag{1}$$

also called the 1-Parameter Logistic (1PL) model, was used to begin analysis. In Equation 1, θ_i represents the ability of student i and β_j represents the difficulty of problem j. Higher values of θ correspond to higher student abilities and higher values of β correspond to harder problems. There is evidence that student abilities and problem difficulties have similar estimates under the 3PL and the Rasch model (Wright, 1995) and so we are not losing much information by starting with the Rasch model. The Rasch model was then extended to the Linear Logistic Test Model (LLTM) which incorporates the transfer model and takes skills into account (Fischer, 1974). In the LLTM, it is assumed that skill requirements for each question combine additively to influence Rasch model question difficulty,

$$\beta_j = \sum_{k=1}^{K} q_{j,k}\alpha_k. \tag{2}$$

Here K is the total number of skills in the transfer model being used and the $q_{j,k}$ are the entries of that transfer model. Thus, β_j is now a linear combination of the skills that appear in problem j. Here α_k represents the difficulty of skill k. Similar to problem difficulties, higher values of α indicate harder skills.

For both the Assistment and MCAS exam dataset, we have discrete observations $\{X_{i,j} : i = 1, 2, ..N, j = 1, 2, ..J\}$. In the Assistment dataset there are $N = 683$ students and $J_A = 354$ questions. The Assistment dataset contains many missing values since no student saw all of the problems. However, these values can be thought of as missing at random and it is not necessary to worry about them in analysis. We have assumed that students are independent of one another and that individual student responses are independent given the student's ability and the problem's difficulty. The model can then be simplified to a Bernoulli trial for each observation,

$$X_{i,j} \sim Bern(P_j(\theta_i)) \tag{3}$$

[3]More information at http://www.doe.mass.edu/mcas/2005/news/03techrpt.pdf.

where $P_j(\theta_i)$ is given above by the Rasch (or LLTM) model above. The complete data likelihood can then be written as

$$P(\underline{X} = \underline{x}) = \prod_{i=1}^{N} \prod_{j:i \text{ saw } j} P_j(\theta_i)^{x_{i,j}} [1 - P_j(\theta_i)]^{1-x_{i,j}}. \tag{4}$$

The same $N = 683$ students are used for analysis of the $J_M = 39$ items from the Spring 2005 MCAS exam. While missing values were possible if students ran out of time or skipped questions, for these students the dataset is complete. Since we are making the same assumptions about the data, the same data likelihood can be applied to this MCAS dataset.

Before continuing it is worth reiterating that we are predicting, given a student's ability and a problem's difficulty, the probability that student i will answer question j correctly. This paper focuses on the β_j values to compare how the Rasch model and LLTM characterize the problems. Ideally, the two models would give similar item difficulties to problems.

To check the fit of the data and transfer model to the IRT models the per problem standardized residuals,

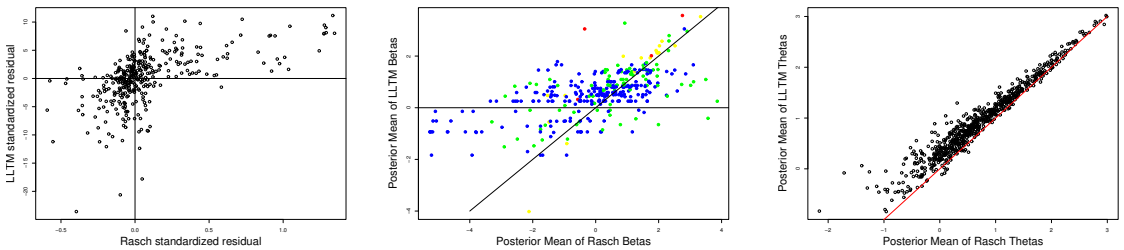
$$r_j = \frac{n_j - E(n_j)}{\sqrt{\hat{var}(n_j)}}, \tag{5}$$

were analyzed. The estimated number of correct answers, $E(n_j)$, was subtracted from the observed number of students who got the question correct, n_j, and the difference was divided by the predicted standard deviation of the number of correct answers. These residuals are actually based on outfit statistics (van der Linden & Hambleton, 1997, Chapter 6), which are based solely on the difference between observed and expected scores.

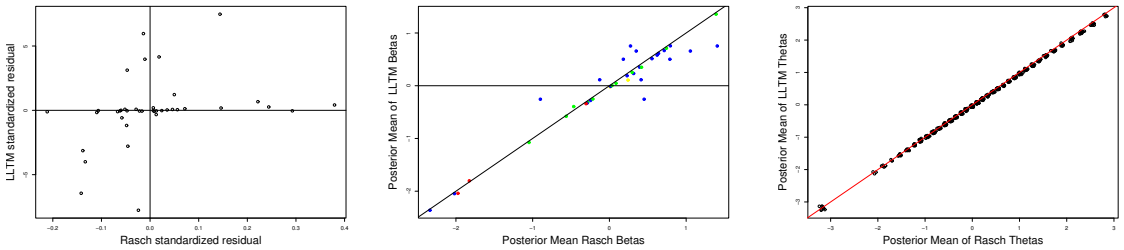
4.1 Results

We estimated the θ_i and β_j values in the Rasch model and the θ_i, β_j, and α_k values in the LLTM, using Markov Chain Monte Carlo methods with the program BUGS (Bayesian Inference Using Gibbs Sampling; Spiegelhalter, Thomas, and Best, 1996). The Rasch model, Equations 1 and 4, was first run on the Assistment data with the priors $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$ and $\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$. The hyperpriors on μ_θ and μ_β were both Normal(0,1E-6) and the hyperpriors on σ_θ^2 and σ_β^2 were both Inverse-Gamma(1,1). The posterior estimate of μ_θ was 0.69 and σ_θ^2 was 0.758 These values were then used as the prior mean and variance of θ in all subsequent simulations as a way of equating. In the LLTM, which constrains the Rasch model by Equation 2, the prior on α was Normal($\mu_\alpha, \sigma_\alpha^2$). The hyperpriors were again Normal(0,1E-6) on the mean and Inverse-Gamma(1,1) on the variance. The same techniques and models were used to obtain posterior estimates for the MCAS dataset.

To determine which problems and skills had non-zero difficulty estimates, equal-tailed posterior 95% credible intervals were used. This was an important check since problems (or skills) with difficulties estimated to be zero are considered middle level and we do not gain much information from these. These



(a) Assistment residuals (b) Assistment Posterior Betas (c) Assistment Posterior Thetas



(d) MCAS residuals (e) MCAS Posterior Betas (f) MCAS Posterior Thetas

Figure 1: Assistment and Spring 2005 MCAS Residuals and Posterior β and θ Estimates

intervals indicated that for the Assistment dataset 64 of the 77 (83.1%) skill parameters in the LLTM model were non-zero. For the 354 problem difficulties, the LLTM gave 334 non-zero estimates and the Rasch model gave 264 non-zero estimates. These results are an indication of a difference between the Rasch model and the LLTM estimates of problem difficulty. In fact, only 175 (of the 354) of the intervals overlap. The other 179 posterior estimates of problem difficulty are significantly different. Thus, for the Assistment data, the constraints of the LLTM are not adequately modeling problem difficulty. For the MCAS data, the LLTM results showed that 20 of the 40 skill parameters were non-zero. In the LLTM 26 of the 39 problem difficulties were non-zero and in the Rasch model 23 of the 39 were non-zero. There was more similarity in posterior estimates as 36 (of the 39) posterior intervals for problem difficulty overlapped.

Figures 1 (a) and (d) show the LLTM versus Rasch residuals, as described in Equation 5. The Rasch model residuals are on the horizontal axis and the LLTM residuals are on the vertical axis. In both cases we see that the Rasch residuals are behaving well, but that the LLTM residuals are much larger than expected. Since these residuals are standardized, we expect them to fall between -2 and 2. This is more evidence that the LLTM is not as adequate as the Rasch model. For the Assistment dataset, the difference in BIC scores between the models is $\sim$ 3100 and thus overwhelmingly favors the Rasch model. When using the MCAS dataset, the difference is $\sim$ 144 and again the Rasch model is favored. The estimated student abilities are shown in Figure 1 (c) and (f). In the Assistment dataset the LLTM estimates higher student abilities than the Rasch model. However, in the MCAS dataset the Rasch and LLTM give practically identical estimates.

a	b	variance of β	Significant $\alpha's$
1	1	2.2660	19
3	3	2.3381	5
10	5	2.2089	5
20	20	2.1584	4
50	50	1.9290	6
100	100	1.6955	10

Table 1: Estimated Posterior Variance of Problem Difficulties and Number of Significant Skill Difficulties for Various Priors in the Random Effects Model

The posterior estimates for of the problem difficulties can be seen in Figures 1 (b) and (e). Rasch model estimates of problem difficulty are on the horizontal axis and LLTM estimates are on the vertical axis. Estimates are color-coded by the number of skills in the problem: blue-1 skill, green-2 skills, yellow-3 skills, and red-4 skills. In general, the problem difficulty increases as the number of skills increases. For the Assistment data the Rasch problem difficulties have less variation than the LLTM problem difficulties. In the plot (b) we see several horizontal lines of dots. These lines indicate problems with the same skill(s) that have the same difficulty estimate from the LLTM, but different Rasch model estimates. This is an effect of the LLTM since we have forced problems with the same skill(s) to have the same difficulty. The same phenomenon occurs within the MCAS dataset but it is not as noticeable since there are only 39 problems.

5 Adding a Random Effect Component

As discussed briefly in Section 4 Figure 1(b) shows horizontal lines. However, there is no reason to believe that every problem with the same set of skill(s) should have the exact same difficulty. One way of modifying this assumption is to add a random effects component, similar to Janssen & De Boeck (2006), to each problem difficulty in the LLTM. The model now states that

$$\mu_j = \sum_{k=1}^{K} q_{j,k}\alpha_k \tag{6}$$

and

$$\beta_j \sim N(\mu_j, \sigma_\beta^2). \tag{7}$$

The prior distribution of σ_β^2 is Inverse-Gamma(a,b) like the previous prior distribution of the variance term of the problem difficulties in the Rasch model. Several different values of a and b have been tried, but in each case the estimated error term is large and overtakes the skill estimates. A much lower percentage of the skill difficulties are significant in these models. In the previous work on Assistment main questions, the estimated posterior variance of the problem difficulties was 1.1080. As can be seen in Table 1, for each of

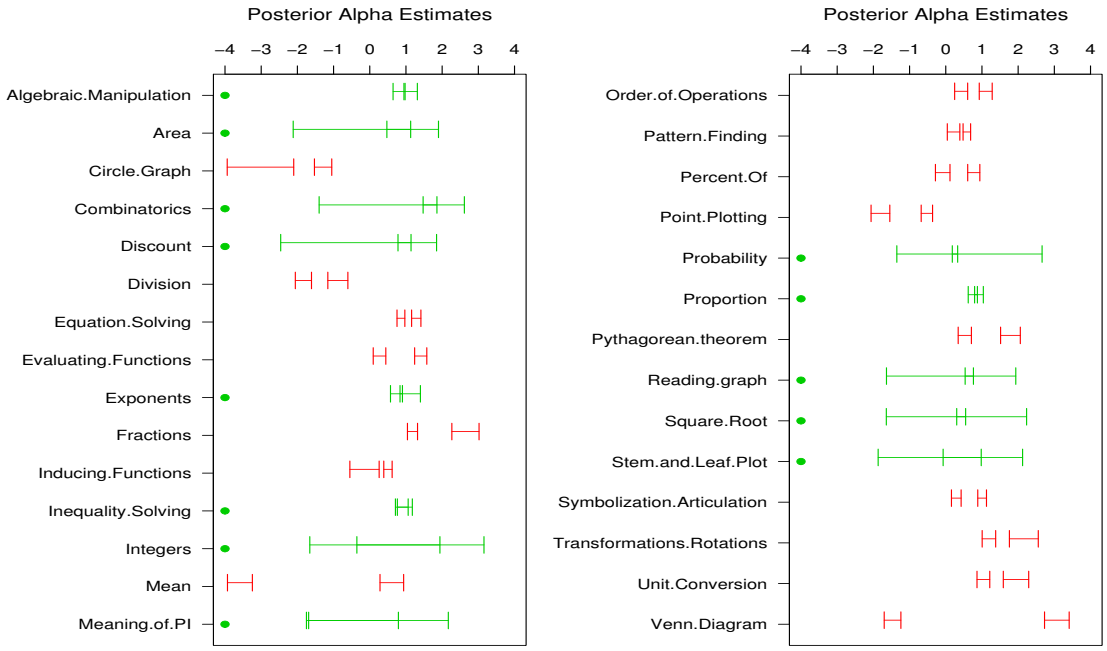


Figure 2: Credible Intervals for the 29 Skills Appearing in both Assistment main questions and the Spring 2005 MCAS Exam.

the prior distributions the posterior estimate of the variance of problem difficulties is much higher. Thus, for our dataset, adding a random effect component does not improve our predictions of problem difficulty.

6 Assessing the Alignment of Skill Difficulty

Although the LLTM gives a poor fit for both the Assistment main questions and the MCAS exam, the skill difficulty estimates from each dataset were compared in search of differences and similarities. There are a total of 29 skills that appear in both the Assistment main questions analyzed and the Spring 2005 MCAS exam. Figure 2 shows the 95% posterior credible intervals for these 29 skills. The intervals in green (with the dots on the skills axis) overlap and the ones in red do not. There are only 13 skills where the intervals overlap. This means that for the other 16 skills, the two models estimated significantly different difficulties. This is a discouraging sign, although it must be taken lightly since we have admitted the neither model is fitting well. Once an adequate skills model is found, an in-depth assessment of any differences in estimated skill difficulties will be informative.

7 Initial Conclusions and Future Plan

An additive skills model for question difficulty, such as the LLTM, does not accurately reflect the range of difficulties of questions, on either Assistment main questions or MCAS test items. For both Assistment main questions and Spring 2005 MCAS exam items, the unrestricted Rasch model has a much better fit than the LLTM. There are at least two possible reasons for this. The first is that the WPI-April-2005 transfer model is not complete: there might be missing skills or the problems may not be correctly coded with skills (for example, there is anecdotal evidence from working with the Assistment System in the classroom, that non-mathematical skills may play a role in question difficulty). Another reason is that this may not be the correct IRT model to use when taking skill knowledge into account. The LLTM model assumes that skills contribute to question difficulty additively. The evidence from this analysis is that an additive skills model for MCAS item test difficulty does not work well. We will explore other IRT models (e.g., conjunctive skills models) for item difficulty and student proficiency in the future.

7.1 Predicting MCAS Exam Scores

As mentioned earlier, significant effort is already underway to generate accurate predictions of MCAS scores from Assistment tutoring data (Anozie & Junker, 2006; Feng et al., 2006; Feng et al. 2006; Pardos et al., 2006; etc.). This paper is a first attempt to build an accurate psychometric model for the tutoring and exam data, in order to improve an apparent lower bound on mean absolute prediction error. Although the LLTM was not a successful model here, we continue to search for a well-fitting psychometric model. Once found, estimated student proficiency from this model can be combined with other Assistment performance metrics to produce an effective prediction function, in an approach that is similar to that of Schofield, Taylor, & Junker (2005).

Acknowledgments

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

This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team including Nathaniel O. Anozie, Brian Junker, 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.

References

Anozie, Nathaniel O and Junker, B.W. (2006) Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. Submitted to AAAI06 Workshop on Educational Data Mining, July 17, 2006, Boston MA.

Barnes, T. M. (2003). *The Q-matrix method of fault-tolerant teaching in knowledge assessment and data mining.* Ph.D. Dissertation, Department of Computer Science, North Carolina State University, Raleigh, NC.

Draney, K. L., Pirolli, P. and Wilson, M. (1995). A measurement model for a complex cognitive skill. Chapter 5 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Embretson, S.E. (1984) A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Farooque, Preeti, and Junker, B.W. (2005) Behavior of Skills within MCAS and Assistment Main Problems.

Feng, M., Heffernan, N.T, Koedinger, K.R. (2006). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. Accepted by the 8th International Conference on Intelligent Tutoring Systems, Taiwan.

Feng, M., Heffernan, N.T, Koedinger, K.R. (2006). Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses, Accepted by the 15th International World Wide Web Conference, Edinburgh, Scotland.

Feng, M., Heffernan, N. T., Mani, M. & Heffernan, C. L. (2006). *Does using finer-grained skill models lead to better predictions of state test scores?* Submitted to the ITS2006 Education Data Mining Workshop, Taiwan ROC, June 26, 2006.

Fischer, G.H. (1974) Einfuhrung in die Theorie psychologischer Tests: Grundlagen und Anwendungen. (Introduction to the Theory of Psychological Tests: Foundations and Applications) Switzerland: Verlag.

Heffernan, N.T., Koedinger, K.R. and 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, US Department of Education. Department of Computer Science at Worcester Polytechnic Institute, Worcester County, Massachusetts. `http://nth.wpi.edu/pubs_and_grants/ Grant_to_IES_with_WPS.pdf`

Janssen, Rianne and De Boeck, Paul. (2006) Technical report, Department of Psychology, University of Leuven, Belgium. A random-effects version of the LLTM.

Junker, B. W. and Sijtsma, K. (2001). Cognitive Assessment models with Few Assumptions, and Connections With Nonparametric Item Response Theory. *Applied Psychological Measurement, 25*, 258-272.

Koedinger, K. R., Corbett, A. T., Ritter, S., & Shapiro, L. J. (2000). Carnegie Learning's Cognitive Tutor: Summary research results. White Paper. Pittsburgh, PA: Carnegie Learning.

Pardos, Z. A., Heffernan, N.T., Anderson, B. & Heffernan, C. L. (2006). *Using fine-grained skill models to fit student performance with Bayesian networks.* Submitted to the ITS2006 Education Data Mining Workshop, Taiwan ROC, June 26, 2006.

Schofield, L., Taylor, L. and Junker, B. W. (2006). The use of cognitive test scores in evaluating black-white wage disparity. Submitted for publication.

van der Linden, Wim J. and Hambleton, Ronald K. (1997; Eds.). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Wright, B.D. (1995) 3PL or Rasch? Rasch Measurement Transactions, 9(1), 408-409.