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

Elizabeth Ayers

Supported in part by IES Training Grant (# R305B040063)

Stage 1: The IRT Models

Per Problem (Rasch Model; Fischer and Molenaar, 1995)

$$P_j(\theta_i) = P(X_{ij} = 1 \mid \beta_j, \theta_i) = 1 / (1 + e^{-(\theta_i - \beta_j)})$$

Additively Per Skill (Linear Logistic Test Model (LLTM); Fischer, 1974)

$$P_j(\theta_i) = P(X_{ij} = 1 \mid \alpha_k, \theta_i) = 1 / (1 + e^{-(\theta_i - \sum_{k=1}^{K} q_{jk}\alpha_k)})$$

In the above equations,

$q_{jk} = 1$ if problem j contains skill k

0 else

Note that the Rasch model corresponds to a unique skill for each problem

Model comparison using Bayesian Information Criterion (BIC; Kass and Raftery, 1995)

- Lower values are better
- Differences larger than 10 denote a significant difference between models

$$BIC = -2 \ln_M + k \log(n)$$

Model	Deviance = $-2\ln_M$	Parameters	BIC
LLTM	56090	79	~56605
Rasch	47640	356	~49963
		Difference in BIC	~6600

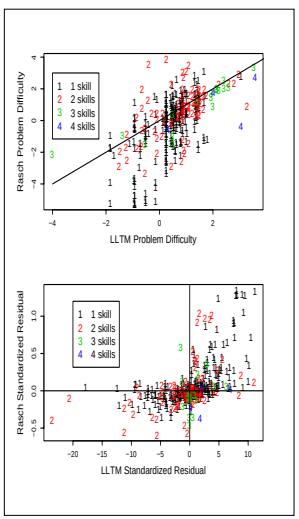
Also looked at the per problem standardized residuals

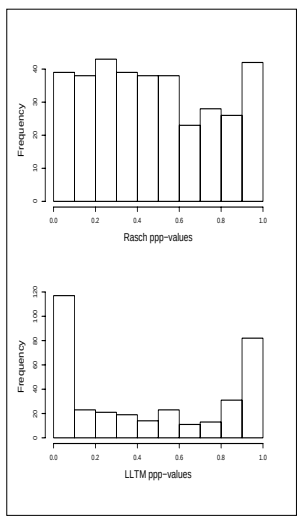
$$r_j = (n_j - E(n_j)) / \sqrt{\text{var}(n_j)}$$

In addition, we calculated the per problem outfit statistics (van der Linden and Hambleton, 1997) To check the fit of each problem, we calculated the posterior predictive p-values (ppp-values) for each observed outfit statistic

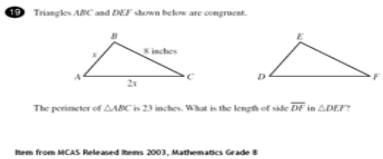
$$p_i \sim 1/M \#\{s: (T_i(x \mid \phi_x) < T_i(x^* \mid \phi_x); s = 1,2,\ldots M\}$$

- ppp-values tend to be conservative (Gelman et al, 1996), but still aggregate around zero if serious misfit for some problems



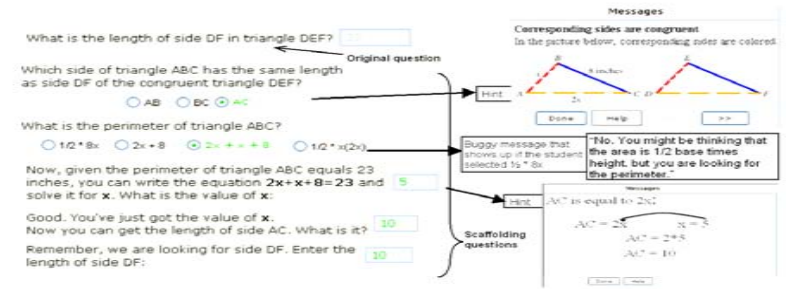


Example of the Assistment System



To the left is Item 19 from the 2003 MCAS exam. Below is the same problem as an Assistment item. The Assistment figure shows two different hints and one buggy message that can occur.

The main question was tagged with three skills: congruence, equation solving, and perimeter. Each scaffold was tagged with only one skill. The first scaffold was tagged with congruence, the second with perimeter, the third with equation solving, and the fourth with congruence.



Stage 2: Calculate Prediction Error Bounds

Assume the MCAS exam (t=1) and the Assistment System (t=2) are parallel tests (Lord and Novick, 1968), we then have

$$X_{i1} = T_i + \varepsilon_{i1} \qquad X_{i1} = T_i + \varepsilon_{i1}$$

The reliability of test *t* is

$$r_t = \sigma^2_T / (\sigma^2_T + \sigma^2_{\varepsilon T})$$

Note that the MSE between the tests can be written as

$$MSE = E[(X_{i1} - X_{i2})^2] = \sigma^2_T ((r_1 + r_2)/(r_1 r_2) - 2)$$

We want to compare models using Mean Absolute Deviation (MAD)

$$MAD = 1/N \sum_{i=1}^{N} |MCAS_i - \text{predicted } MCAS_i|$$

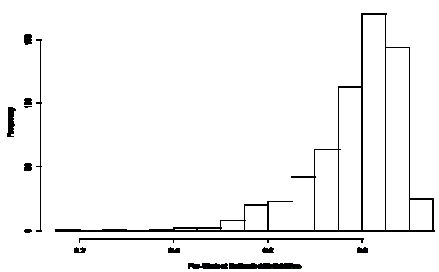
Note that we can bound MAD with MSE

$$1/(\max_i |MCAS_i - \text{Pred}_i|) \, MSE \le MAD \le MSE$$

Since each student completed a unique set of Assistment main questions, we calculated a per-student Assistment reliability using Cronbach's alpha coefficient (Cronbach, 1951). The figure below shows the histogram of the calculated reliabilities.

Using $r_1 = 0.9190$ and $\sigma^2_T = 130.86$ along with the median student reliability of 0.8180 yields the following prediction error bound

$$1.053 \le MAD \le 6.529$$



Stage 3: Predicting MCAS Exam Scores

Combine student proficiency estimate from the Rasch model with tutor metrics to predict MCAS exam scores

$$MCAS_i = \lambda_0 + \lambda_1 \text{ Rasch Student Proficiency} + \lambda_2 \text{ Percent Correct on Scaffolds} + \lambda_3 \text{ Seconds Spent on Incorrect Scaffolds} + \ldots + \varepsilon_i$$

Model	Variables	# of Variables	CV MAD	CV RMSE	Notes
Model 1	Percent correct on main questions	1	7.18	8.65	
Model 2	Rasch Student Proficiency	1	5.90	7.18	
Model 3	Percent Correct on Main questions and 4 other tutor metrics	35	5.46	7.00	Uses multiple monthly summaries
Model 4	Rasch student proficiency and same 4 tutor metrics as Model 3	5	5.39	6.56	Uses only year-end aggregates
Model 5	Rasch student proficiency and 5 tutor metrics (one overlap with 3, 4)	6	5.24	6.46	Optimized for student proficiency

Variables

Variable Name	Model	Definition
Student Proficiency	2, 4, 5	IRT estimate of student proficiency
PctCorMain	1, 3	Percent of correctly answered main questions
PctCorScaf	3, 4	Percent 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 incorrect main questions) / number of main questions attempted
SecCorScaf	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	Percent of time on main questions spent on incorrect main questions

Overall Conclusions

Stage 1: The less parsimonious Rasch model provides a better fit

Stage 2: Calculations yield the prediction error bounds

$$1.053 \le MAD \le 6.529$$

Stage 3: Using a Rasch estimate of student ability with tutor metrics improves the predictions of MCAS exam scores

Advisor: Brian Junker

This work would not have been possible without the assistance of the2004-2005 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.