Assessing Students' Performance Longitudinally: Item Difficulty Parameter vs. Skill Learning Tracking

Mingyu Feng, Neil T. Heffernan Worcester Polytechnic Institute {mfeng|nth}@wpi.edu

Objective

Most large standardized tests (like the math-subtest of the Graduate Record Examination (GRE)) analyzed with Item Response Theory are "unidimensional" in that they are analyzed as if all the questions are tapping a single underlying knowledge component (i.e., skill). However, cognitive scientists such as Anderson & Lebiere (1998), believe that students are learning individual skills, and might learn one skill but not another. Among the reasons that psychometricians analyze large scale tests in a unidimensional manner is that students' performance on different skills are usually highly correlated, even if there is no necessary prerequisites relationship between these skills. Another reason is that students usually do a small number of items in a given setting (e.g. 39 items for the 8th grade math Massachusetts Comprehensive Assessment System test). We are engaged in an effort to investigate if we can do a better job of predicting a large scale test (MCAS) by modeling individual skills in different grain-sized skill models than by using item difficulty parameters induced from traditional Item Response Theory models, on which computer adaptive testing relies. We consider 2 different skill models¹, one has 5 skills we call the "WPI-5", and the other is our most fine-grained model has 78 skills we call the "WPI-78". In both cases, a skill model is a matrix that relates questions to the skills needed to solve the problem. The measure of model performance is the accuracy of the predicted MCAS test score based on the assessed skills of the students.

Given that the WPI-78 composed of 78 skills, people might worry about that we were overfitting our data by fitting a model with so many free parameters. However, we were not evaluating the effectiveness of the skill models over the same online ASSISTment data based on which the models will be constructed. Instead, we used totally different data (from the external, paper-and-pencil based state test) as the testing set. Hence, we argue that overfitting would not be a problem in our approach.

Modeling student responses data from intelligent tutoring systems has a long history (Corbett, Anderson, & O'Brien, 1995; Draney, Pirolli, & Wilson, 1995). Corbett and Anderson did show that they could get better fitting models to predict student performance in LISP programming by tracking individual production but their system never asked questions that were tagged with more than one production, which is the sort of data we have (described below). Our collaborators (Ayers and Junker, 2006) are engaged trying to allow multi-mapping2 using a version of the WPI-78 but report their LLTM model does not fit well. Anozie & Junker (2006), are looking at this same data set,

¹ What we refer to as a "skill model" is referred to as "Q-Matrix" by some AI researchers (Barnes, 2005) and psychometricians (Tatsuoka, 1990); and in Hao, Koedinger & Junker (2005), they used the term "cognitive model", while Croteau, Heffernan & Koedinger (2004) used the term "transfer model".

² A "multi-mapping" skill model, in contrast to a "single-mapping" or a "non-multi-mapping" model, allows one item to be tagged with more than one skills.

also trying to predict the same state test scores we will describe below, but they are not using skills at all, and in that since, their method is unidimensional, in one sense representing the more traditional psychometric approach.

IRT now underlies several major tests. Computerized adaptive testing (CAT), in particular, relies on IRT. In CAT, examinees receive items that are optimally selected to measure their potential. IRT principles are involved in both selecting the most appropriate items and equating scores across different subsets of items. IRT now contains a large family of models. The simplest model is the Rasch model, also known as the one-parameter logistic model (1PL). For the model, the dependent variable is the dichotomous response for a particular person to a specified item. The independent variables are the person's trait score, θ_s , and the item's difficulty level, β_i .

Though the Rasch model itself can be used to estimate the probability of the success response on specified items, in order to compare the effectiveness of difficulty parameter (β) with the skill learning tracking technique on predicting students' performance, our approach is not using pure Rasch model. Instead we introduced either β or skill as a covariate in our mixed-effects logistic regression models and then examined to see which model leads to more accurate prediction. It turned out that in our case we can do a better job predicting students' MCAS test scores by doing skill learning tracking than by using the difficulty parameters.

Source

The Massachusetts Comprehensive Assessment System (MCAS)

MCAS is a Massachusetts state administered standardized test that produces tests for English, math, science and social studies for grades 3 to 10. We focused on only 8th grade mathematics. Our work is related to the MCAS in two ways. First we built out content based upon released items. Secondly, we evaluate our models using the 8th grade 2005 test, which we will refer to as the state test. Predicting students' scores on this test will be our gauge of model performance. The state test consists of 5 open response, 4 short answer and 30 multiple choice (out of 4) questions. Only the multiple choice and short answer questions are used in our prediction with regard to the fact that currently open response questions are not supported in our system. This makes a full score of 34 points with one point earned for a correct response on an item. For the students in our data set, the mean score out of 34 points was 17.9 (standard deviation=7.1).

The ASSISTment System

The ASSISTment system is an online tutoring system that is about 2 years old. In the 2004-2005 school year some 600+ students used the system about every two weeks. 8 math teachers from two schools would bring their students to the computer lab, at which time students would be presented with randomly selected MCAS test items. In Massachusetts, the state department of education has released 8 years (1998-2005) worth of MCAS test items, over 300 items, which we have turned into ASSISTments by adding "tutoring". If students got the item correct they were given a new one. If they got it wrong, they were provided with a small "tutoring" session where they were forced to answer a few questions that broke the problem down into steps. The key feature of

ASSISTments is that they provide instructional assistance while assessing students. Razzaq & Heffernan (2006) addressed student learning due to the instructional assistance, while this paper is focused on skill model evaluation by assessing students' performance on a state test.

Each ASSISTment consists of an original question and a list of scaffolding questions. The original question usually has the same text as in MCAS test while the scaffolding questions were created by our content experts to coach students who fail to answer the original question. An ASSISTment that was built for item 19 of the 2003 MCAS is shown in Figure 1. In particular, Figure 1 shows the state of the interface when the student is partly done with the problem. The first scaffolding question appears only if the student gets the item wrong. We see that the student typed "23" (which happened to be the most common wrong answer for this item from the data collected). After an error, students are not allowed to try the item further, but instead must then answer a sequence of scaffolding questions (or "scaffolds") presented one at a time. Students work through scaffolding questions. the

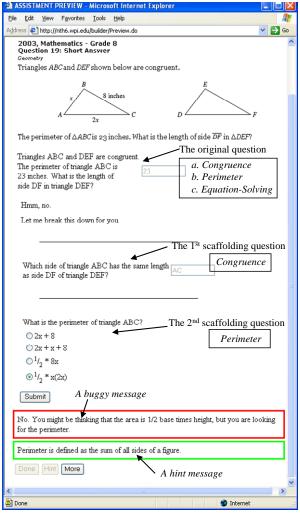


Figure 1. An ASSISTment, showing 2 scaffolding questions, one buggy message and a hint message that can occur at different points.

possibly with hints, until they eventually get the problem correct. If the student presses the hint button while on the first scaffold, the first hint is displayed, which would be the definition of congruence in this example. If the student hits the hint button again, the second hint appears which describes how to apply congruence to this problem. If the student asks for another hint, the answer is given. Once the student gets the first scaffolding question correct (by typing "AC"), the second scaffolding question appears. Buggy messages will show up if the student types in a wrong answer as expected by the author. Figure 1 shows a buggy messages that appeared after the student clicked on " $\frac{1}{2}$ *x(2x)" suggesting he might be thinking about area. Once the student gets this question correct he will be asked to solve 2x+x+8=23 for 5, which is a scaffolding question that is focused on equation-solving. So if a student got the original question wrong, what skills should be blamed? This example is meant to show that the ASSISTment system has a better chance of showing the utility of fine-grained skill modeling due to the fact that we can ask scaffolding questions that will be able to tell if the student got the question wrong because they did not know congruence versus not

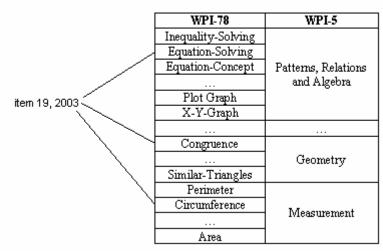


Figure 2: Hierarchical relationship among skill models

knowing perimeter, versus not being able to set up and solve the equation. As a matter of logging, the student is only marked as getting the item correct if they answered the questions correctly before asking for any hints or encountering scaffolding.

Skill Models

In April, 2005, we staged a 7 hour long "coding session", where our subjectmatter expert, Cristina Heffernan, with the assistance of the 2nd author set out to make up skills and tag all of the existing 8th grade MCAS items with these skills.³ There were about 300 released test item for us to code. Because we wanted to be able to track learning between items, we wanted to come up with a number of skills that were somewhat fine-grained but not too fine-grained such that each item had a different skill. We therefore imposed upon our subject-matter expert that no one item would be tagged with more than 3 skills. She was free to make up whatever skills she thought appropriate. We printed 3 copies of each item so that each item could show up in different piles, where each pile represented a skill. Although we have English names for the skills, those names are just a handy tag; the real meaning of a skill must be divined by the questions with which it is associated. The name of the skill served no-purpose in our computerized analysis. When the coding session was over, we had 6 foot-long tables covered with 106 piles of items. We wound up with about 106 skills, but not every skill that was created was eventually involved in the data source used by this work so we call this model the WPI- 78^4 . To create the coarse-grained model, the WPI-5, we used the fine-grained model to guide us. We decided to use the same 5 categories that both the National Council of Teachers of Mathematics uses, as well as the Massachusetts Department of Education. These categories are named 1) "Patterns, Relations and Algebra", 2) "Geometry", 3) "Data Analysis, Statistics and Probability", 4) "Number Sense and Operations" and 5) "Measurement". Then the WPI-5 model is derived from the WPI-78 by nesting a group of fine-grained skills into a single category. The Massachusetts Department of Education

³ We hand-coded the skills in this work. Though, we believe it is possible to use an automatic technique such as LFA (Hao, Koedinger & Junker, 2005) or Q-matrices (Barnes, 2005) for topic construction.

⁴ In Pardos, Heffernan, Anderson & Heffernan (2006) we called this model the WPI-106 because they used a data set that included additional items.

actually tags each item with exactly one of the 5 categories, but our mapping was not the same as the states'. Furthermore, we allowed multi-mapping, i.e. allow an item to be tagged with more than one skill. An interesting piece of future work would be to compare our fit with the classification that the state uses. After the students had taken the 2005 state test, the state released the items in that test, and we had our subject-matter expert tag up these items in WPI-5 and WPI-78. Figure 2 shows the hierarchal nature of the relationship between WPI-78 and WPI-5. The first column lists 10 of the 78 skills in the WPI-78 skill model. In the second column we see how the 5 skills in WPI-78 are nested inside of "Patterns, Relations and Algebra", which itself is one piece of the 5 skills that comprise the WPI-5 skill model. Consider the item 19 from 2003 MCAS test (See Figure 1). In the WPI-78 skill model, the first scaffolding question is tagged with "congruence", the second tagged with "perimeter", the third tagged with "equation-solving". In the WPI-5, the questions were therefore tagged correspondingly with "Geometry", "Measurement" and "Patterns, Relations and Algebra"

Data Source

We collected online data of 497⁵ students who used our system from Sep. 17, 2004 through May 16, 2005 for on average 7.3 days (one period per day). All these students have worked on the system for at least 6 days. The item-level state test report is available for all these students so that we were able to construct our predictive models on these students' data and evaluate the accuracy on state test score prediction. The original data set, corresponding to students' raw performance, includes both responses to original questions and to scaffolding questions. It contains about 138 thousand data points, among which around 43 thousand come from original questions. On average, each student answered 87 MCAS (original) questions and 189 scaffolding questions.

We then created different versions of the data set. When skill models being used, the data is organized in the way that there can be one or multiple rows for every student response to each single question depending on what's the skill model we are interested in and how many skills the question is "tagged" with in that particular skill model. For instance, suppose a question is tagged with 2 skills in a model, then for each response made to the question there would be 2 rows in the data set, with skill names listed in a separate column. Students' exact answers are not included. Instead, we use a binary column to represent whether the student answered the specified item correctly or not. No matter what the input type of the item is (multiple-choice or text-input), a "1" indicates a correct response while a "0" means a wrong answer was given. Additionally, a column is associated with each response, indicating the number of months elapsed since September 17, 2004 till the time when the response was made. Thus the number of months elapsed for a response made on September 17th will be zero, and the number will 1 for a response made at October 17th, 2004, and so on. This gives us a longitudinal, binary response data set across the school year. In another version of the data set, we organized the data in a similar way, but instead of adding in related skills, we appended the item difficulty parameter in a separate column and of course there were no duplicate rows as described above because of multi-tagged skills.

Table 2 displays 12 rows of the raw data for one student (system ID = 950) who finished the item 19 (shown in Figure 1) and item 27 (shown in Figure 2) on two different

⁵ The amount of data is limited by the maximum memory allowed by the open source statistical package (R) we used.

RowID	StudentID	State Test ID	ItemID	WPI-78 skills	Original?	Response	Month Elapsed
1	950	2003-#19	326	Congruence	Y	0	1.32
2	950	2003-#19	326	Perimeter	Y	0	1.32
3	950	2003-#19	326	Equation-Solving	Y	0	1.32
4	950	2003-#19	327	Congruence	Ν	0	1.32
5	950	2003-#19	328	Perimeter	Ν	1	1.32
6	950	2003-#19	329	Equation-Solving	Ν	0	1.32
7	950	2003-#19	330	Equation-Solving	Ν	0	1.32
9	950	1999-#27	1183	Perimeter	Y	0	2.94
10	950	1999-#27	1183	Area	Y	0	2.94
11	950	1999-#27	1184	Perimeter	Ν	1	2.94
12	950	1999-#27	1185	Area	Ν	1	2.94

Table 1. Sample Raw Data (Skill Tracking Version)

days. Here the WPI-78 skill model was used as an example. The first 7 rows represent the student' response on item 19 (with original item ID^6 being 326) and the rest 6 rows show his response on item 27 (with original item ID being 1183). We can see that since the original question of item 19 was tagged with 3 skills "Congruence", "Perimeter" and "Equation-Solving", the student's response was duplicated in row 1 - 3 and so does the original question of item 27 as in row 9 and row 10. For both items, the student answered the original questions wrong (indicated by "0" in the response column of row 1-3 and row 9-10) and thus was presented the scaffolding questions. The student did not do very well on the first item. He only gave a correct answer to the second scaffolding question (indicated by "1" in the response column of row 5), but failed to answer all the other scaffolding questions. On contrast, for item 27, though not getting the original question right on the first shot, the student went through all three scaffolding questions correctly.

Methods

Mixed-effects Logistic Regression Modeling

Mixed-effects logistic regression model is a very popular and widely accepted choice for analysis of dichotomous data (Snijders & Bosker,1999; Hedeker & Gibbons). It describes the relationship between a binary or dichotomous outcome and a set of explanatory variables. In this work, we adopted this modeling approach and fitted on our longitudinal, binary response data, using *Time*, and either *Skills* or *item difficulty parameter* as predictors to predict the probability that a student will correctly answer an item of certain difficulty and tagged with particular skills at certain time. And our aim is to tell which predictor helps more to construct a better-fitted model and thus better estimate students' score on the state test. Such a model is often referred to as "longitudinal model" (Singer & Willett, 2003) since *Time* is introduced as a predictor of the response variable, which allows us to investigate change over time. After the model was constructed, the two learning parameters "intercept" (indicating initial status) and "slope" (representing learning rate) were calculated for each skill and for each individual student. Given these, we thus can apply the model on the items in the state test to estimate students' response to each of them.

Getting Item Difficult Parameters

⁶ The "itemID" is a number that we used internally in the system to uniquely identify a question. It is displayed only for the purpose of interpreting the data.

To get the β 's for the ASSISTments, we were using 2005-2006 ASSISTment data for the same group of items but done by a different group of 2702 students from the same district as the 497 students in our data, assuming students from different years are of the same knowledge level. After training up the Rasch model, we extracted the β 's for all the items and observed that the values of β center around zero and range⁷ from -2.37 to 2.69. Then we added a new column in our data (See the sample data in Table 1.) putting in the corresponding β for the particular item in the each row. Now the data is ready to be used to train mixed-effects logistic regression models with β as a covariate. The similar approach was followed to get the β 's for the state test items. The item level response data of 1117 8th graders from Worcester who have not gotten involved in the ASSISTment system was utilized to train the Rasch model and we observed that the β 's of the 34 state test items range from -2.22 to 1.60.

Measuring Model Performance

The accuracy of the predicted MCAS test score was used to evaluate different approaches. Specifically, we trained 3 mixed-effects regression models. All these are longitudinal models with *Time* being used as one predictor and the dependent variables in all models are the same, that is, the probability that a student will respond correctly to an item at a time. In addition to *Time*, the first model, **Model-Beta**, includes item difficulty parameter as another predictor, while in the second model, **Model-WPI-5**, skills in the WPI-5 is used as a predictor; and the third model, **Model-WPI-78**, used skills in the WPI-78 as the other predictor other than *Time*. These models were constructed based on the ASSISTment online data and applied to MCAS test items to give an assessment of students' scores.

To predict a student's total test score, we will first find the fractional score the student can get on each individual item in the MCAS test and then sum the "item-score" up to acquire a total score for the test. So how did we come up with a prediction of their item-score? The first thing we did is identifying what are the skills associated with the item in both skill models and what is the item difficulty level of each item in the state test, depending which model was used. Then, given a student's learning parameters, for any particular item in the state test, we can calculate the probability of positive response from the student. In the case that an item was tagged with more than one skill (i.e., when WPI-5 and WPI-78 was used as the skill model), we picked the lowest probability among all the skills that apply to the item⁸ for that student⁹. In our approach, a student's probability of correct response for an item was used directly as the fractional score to be awarded on that item for the student. Therefore, once we obtained the probability of correct response for all the items, we sumed them up to produce the total points awarded. For example, if the probability of an item marked with Geometry is 0.6, then 0.6 points are added to the

⁷ A higher value of β indicates the item is relatively easy while a lower one suggests a relative harder item.

⁸ We admit that there are other approaches dealing with multi-mapped items. For instance, one way can be taking into consideration the conjunctive relationship among the skills and "somehow" combining the probabilities together to produce a "final" probability for the item. Using Bayesian Networks is also a reasonable way to deal with this situation and our colleague Pardos, Hefernan, Anderson and Heffernan (2006) use this approach and seem to getting similar results that fine grained models enable better predictive models.

⁹ We consider the skill that had the lowest probability of correct response in our model the hardest skill for a student.

sum to produce the points awarded. This sum of these points is what we use as our prediction of their state test score¹⁰.

To compare prediction models we use the mean absolute deviation (MAD) (equal to *average*(|MCAS_ predicted MCAS|)). We also report a normalized metric, the percent error, by dividing the MAD score by 34 to reflect the fact that the MAD is out of a full score of 34 points.

Results

As shown in Table 2, Model-WPI-78 shows the highest accuracy among all the three models. Model-Beta, which used *Time* and *item difficulty parameter* to predict student performance, does not do as well as the two skill learning tracking models, Model-WPI-5 and Model-WPI-78. The paired t-test that compares the absolute difference between real scores and the predicted scores shows that the accuracy of prediction of Model-WPI-5 and Model-WPI-78 are both statistically better than that of Model-Beta (p < 0.001 for both comparisons). On top of that, we did a similar paired t-test to compare the performance of the models Model-WPI-5 and Model-WPI-78 and found that as a finer-grained skill model, WPI-78 did a significant better job than WPI-5 on tracking student learning and reach an error rate 11.97% as we can see in Table 2.

At the first blush, an error rate of 11.97% seems hardly dramatic. So we asked ourselves: Can we do better? Should we be dissatisfied unless we can get the MAD of zero? We want to investigate what a reasonable comparison should be. Ideally, we wanted to see how good one MCAS test was at predicting another MCAS test. We could not hope to do better than that. We did not have access to data for a group of kids that took two different version of the MCAS test to measure this, but we could estimate this by taking our students scores on their real MCAS test, and randomly spiting the test in half, and then using their score on the first half to predict the second half. We excluded open response questions from the 39 items in MCAS 2005 test and kept the remaining 34 multiple-choice and short answer questions with regard to the fact that open response questions are not supported in the ASSISTment system currently. Then the 34 items were randomly split into two halves and student performance on one half was used to predict their performance on the other half. This process was repeated 5 times. On average, we got MAD of 1.89, which is about 11% of the full score (17 points with one point for each item).

¹⁰ We think it might be useful to discuss your model from a more qualitative point of view. Is it the case that if you tag an item with more skills, does that mean our model would predict that the item is harder? The answer is not , in that sense that if you tagged a bunch of item with a easy skill (i.e., one easier then what the item was currently tagged with), that would not change our models prediction at all. This makes qualitative sense, in that we believe the probability of getting a question correct is given by the probability of getting correct the most difficult skill associated with that question.

	Real	Predicted MCAS score			Abs(real score – predicted score)					
Students	MCAS	Model-	Model-	Model-	Model-	Model-	Model-			
	score	Beta	WPI-5	WPI-78	Beta	WPI-5	WPI-78			
Tom	22	20.91	19.86	17.28	1.09	2.14	3.72			
Dick	26	24.15	23.76	20.96	1.85	2.24	5.04			
Harry	25	19.08	17.76	16.21	5.92	7.24	7.79			
Mary	25	20.44	19.18	18.38	4.56	5.82	5.62			
492 rows omitted										
Lisa	9	17.04	17.35	15.87	8.04	8.35	6.87			
				MAD	4.63	4.47	4.07			
				%Error	13.63%	13.15%	11.97%			
				(MAD/34)	15.05%	15.15%	11.97 70			

Table 2 . Comparing models

Conclusion & Educational importance

It appears that we have found evidence that shows skill learning tracking can better predict MCAS score than simply using item difficulty parameter and fine-grained models did even better than coarse-grained model. The result is consistent with our previous work (Feng et al., 2006). And we believe that the ASSISTment system can be an even better predictor of the state test scores because of this work. Of course, teachers want reports by skills, and this is evidence we have saying that our skill mappings are "good" (We make no claim that the WPI-78 is optimal.) And to predict state test score by doing skill tracking is also more practical than based on item difficulty parameter which won't be available before the test. Now that we are getting reliable models showing the value of these skill models, we will consider using these models in selecting the next best-problem to present a student with. As part of the future work, we will get our data ready to be shared with other scholars.

Acknowledgements

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 of the opinions in this article are those of the authors, and not those of any of the funders.

Reference

Anderson, J. R. & Lebiere, C. (1998). The Atomic Components of Thought. LEA.

Anozie N., & Junker B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In Beck, J., Aimeur, E., & Barnes, T. (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 1-6. Technical Report WS-06-05.

- Ayers E., & Junker B. W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In Beck, J., Aimeur, E., & Barnes, T. (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 14-20. Technical Report WS-06-05.
- Barnes, T., (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Technical Report WS-05-02. ISBN 978-1-57735-238-9.
- 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, *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.
- Croteau, E., Heffernan, N. T. & Koedinger, K. R. (2004). Why Are Algebra Word Problems Difficult? Using Tutorial Log Files and the Power Law of Learning to Select the Best Fitting Cognitive Model. In J.C. Lester, R.M. Vicari, & F. Parguacu (Eds.) *Proceedings of the 7th International Conference on Intelligent Tutoring Systems.* Berlin: Springer-Verlag. pp. 240-250.
- Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, New Jersey.
- 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 8th International Conference on Intelligent Tutoring Systems. Springer-Verlag: Berlin. pp. 31-40. 2006.
- Feng, M., Heffernan, N. T., Mani, M., & Heffernan, C. (2006). Using Mixed-Effects Modeling to Compare Different Grain-Sized Skill Models. In Beck, J., Aimeur, E., & Barnes, T. (Eds). *Educational Data Mining: Papers from the AAAI Workshop*. Menlo Park, CA: AAAI Press. pp. 57-66. Technical Report WS-06-05. ISBN 978-1-57735-287-7.
- Hao C., Koedinger K., and Junker B. (2005). Automating Cognitive Model Improvement by A*Search and Logistic Regression. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Technical Report WS-05-02. ISBN 978-1-57735-238-9.
- Hedeker, D. & Gibbons, Robert. D. (2006). "Longitudinal Data Analysis": "Mixed-Effects Regression Models for Binary Outcomes" (chapter 9).
- Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan C. (2006). Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. Workshop on Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems, Taiwan, 2006.
- Razzaq, L., Heffernan, N. T. (2006). Scaffolding vs. Hints in the Assistment System. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Berlin: Springer-Verlag. pp. 635-644. 2006.
- Singer, J. D. & Willett, J. B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Occurrence. Oxford University Press, New York.
- Snijders, Tom A. B., and Bosker, Roel J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London etc.: Sage Publishers, 1999.
- Tatsuoka, K.K. (1990). Toward an Integration of Item Response Theory and Cognitive Error Diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.
- van der Linden, W. J. and Hambleton, R. K. (eds.) (1997) Handbook of Modern Item Response Theory. New York: Springer Verlag.