Back to the future: a non-automated method of constructing transfer models

Mingyu Feng and Joseph Beck
{mfeng, josephbeck}@wpi.edu
Computer Science Department, Worcester Polytechnic Institute

Abstract. Representing domain knowledge is important for constructing educational software, and automated approaches have been proposed to construct and refine such models. In this paper, instead of applying automated and computationally intensive approaches, we simply start with existing hand-constructed transfer models at various levels of granularity and use them as a lens to examine student learning. Specifically, we are interested in seeing whether we can evaluate schools by examining the grain-size at which its students are best represented. Also, we are curious about whether different types of students are best represented by different transfer models. We found that better schools and stronger students are best represented by models with a fewer number of skills. Weaker students and schools are best represented, for our data, by models that allow no transfer of knowledge in between skills. Perhaps surprisingly, to accurately predict the level at which a student represents knowledge it is sufficient to know his standardized test score rather than indicators of socio economic status or his school.

1 Introduction

The topic of representing domain knowledge is fundamental in the construction of intelligent tutoring systems (ITS). This representation is important not only because it denotes the language used in constructing the tutor (e.g. the level at which to construct hints), but also because it makes claims about the level at which students represent knowledge and transfer it between problems. For this reason, such models are sometimes called transfer models [7].

Given the importance of transfer models, it is not surprising that their construction has been a major focus in the educational data mining (EDM) community. For example, Barnes has done considerable work with trying to induce transfer models, in this work called q-matrices [4], from data [1, 2]. Winters [16] has compared a variety of statistical approaches for constructing transfer models, including cluster methods such as k-means and dimensionality reduction such as non-negative matrix factorization. One common thread of this work is that it produces models that are typically more compact than those created by experts. This difference is both a source of strength (perhaps students learn differently than experts believe?) and a source of weakness (if the models are less understandable or make it harder to represent pedagogical knowledge why should we use them?). Although it would be an expensive undertaking, we are unaware of a controlled study showing that a tutor using automatically constructed model provides superior teaching compared to a tutor using to hand-constructed transfer model (or vice versa). Rather than inferring a transfer model from scratch, there is a hybrid approach called learning factors analysis (LFA) [9]. This technique starts with a transfer model, typically built by hand, and computationally tries various modifications to the model to better align it with student performance data (e.g. see [5, 6]). Although LFA is intuitively appealing, it has not demonstrated any dramatic improvements in model fit. Although its potential to shed light on scientific questions, such as the level of knowledge that learners use to represent written words [11], is a great benefit, it is unfortunate that modifying hand-created models does not result in substantially stronger[1] models.

Rather than trying to invent another complex and computationally intensive technique, we take an alternate view of the problem. We know from prior research that students of differing proficiency have somewhat different representations of the domain, with more skilled learners having a more compact (i.e. coarser) representation [11]. We also know that different tutorial interventions influence the representation that learners acquire, with better interventions causing learners to develop a more compact representation [10]. A common fallacy is the belief that finer-grain models will fit learner data better, or at least will fit better given sufficient training data, since they are able to represent subtler distinctions in the domain. This belief is incorrect since fine-grained models not only make subtle distinctions in skills, they (typically) also assert that skills are independent of each other. So practice in one skill does not help with another. If learners are able to transfer knowledge amongst skills, a coarser-grain model will better fit performance data. Given these results, perhaps it makes more sense to skip over automated techniques and simply start with transfer models at various levels of granularity and use them as a lens to examine student learning. In this way, we can still do interesting science with our large datasets but do not have to focus on complex machinery that might not be that helpful.

The goal of this paper is a case study in hand-constructed models of various grain sizes in interpreting data collected from an ITS. Specifically, we are interested in whether models of different granularity better fit distinct subgroups, and, consequently, whether we can use this approach to evaluate schools by examining the grain-size at which their students are best represented. Given two schools where one is better predicted by a coarser transfer model, that school is probably the better one. This approach is different than simply looking at which school has the highest test score performance. If a weaker school changes its curriculum and its students have a better mental model of the domain and are transferring better, they might still lag a stronger school in raw knowledge and consequently in test scores. This approach can potentially detect such schools. We can validate this hypothesis in two ways. First, we have an idea of the quality of the schools we are evaluating (although the person interpreting the data did not). Second, instead of partitioning students by school, we can use their state assessment test score and partition them by math proficiency. If we see a trend for stronger students, it is reasonable to believe it applies to stronger schools.

The advantage of this approach is that it is easy. Also, if one transfer model does a better job at a particular school, since that model is expert-constructed it should not be (any more) difficult (than usual) to construct tutorial content for the model; whereas automated models might not fit educators' understanding of the domain. Also, since there are a

[1] By "substantially stronger" we do not mean statistically reliably different. We acknowledge there have been changes in Bayesian Information Criterion (BIC) scores that correspond to reliable improvements, but it would be difficult to distinguish such a tutor built with a revised transfer model from the one built with the original.

limited number of grain sizes for our model, there is a definite limit on the amount of content creation that is required. For this research, we use data collected as part of the ASSISTment project (www.assistment.org) as our testbed.

2 The ASSISTment system

The ASSISTment system [13] is a web-based system that presents math problems to students who range from approximately 12 to 16 year-olds. When a student has trouble solving a problem, the system usually provides instructional assistance to lead the student through by breaking the problem into scaffolding steps, or displaying hint messages on the screen, upon student request. Each ASSISTment question consists of an *original question* and a list of *scaffolding questions*. The original question usually has the same text as found in the Massachusetts Comprehensive Assessment System (MCAS) test while the scaffolding questions were created through breaking the original question down to the individual steps by our content experts. A student is initially presented a question that usually has several skills needed to solve it correctly. If the student gets the question correct he can move on to next question, otherwise he is forced to go through a sequence of scaffolding questions (or scaffolds). Students work through the scaffolding questions, possibly with hints and buggy messages, until they eventually get the problem solved. Student actions and tutorial responses are time-stamped and logged into our database.

3 Methods

3.1 Construction of different grain sized transfer models

A fine grained model was constructed during a seven hour long “coding session” in 2005 at WPI where our subject-matter expert and the ASSISTment project director created a set of skills and used those skills to tag all of the existing 8th grade MCAS items. They imposed the limit that no one item would be tagged with more than three skills. Thus, many of our ASSISTment System questions had three scaffolding questions; we wanted the fine grainedness of the modeling to match the fine grainedness of the scaffolding. During the “coding session”, the subject-matter expert reviewed the problems and conducted a cognitive task analysis to identify what knowledge was needed to perform each task. When the coding session was over, we wound up with about a model of 106 skills, called the WPI-106 model. To create the coarse-grained models, we used the fine-grained model to guide us. We decided to use the same five broad strands that were used by the Massachusetts Department of Education to tag each MCAS item with exactly one strand. Since our mapping was inferred from the WPI-106, it was not the same as the state’s mapping. Therefore, it was named the WPI-5. Furthermore, we allowed multi-mapping, i.e., allowing an item to be tagged with more than one skill. Similarly, we adopted the name of the 39 learning standards (nested inside the five strands) in the Massachusetts Curriculum Framework, associated each skill in the WPI-106 to one of the learning standards, and thus we created the model WPI-39. This process is illustrated in Table 1. After the students had taken the state tests, the state released the items in that test, and our subject-matter expert tagged up these items in all the transfer models.

The first column in Table 1 lists eight of the 106 skills in the WPI-106 model. For instance, *equation-solving* is associated with problems involving setting up an equation and solving it; while *equation-concept* is related to problems that have to do with equations in which students do not actually have to solve them. The two skills are nested inside of "Patterns, Relations and Algebra" in the third column which itself is one piece of the five skills that comprises the WPI-5 transfer model. The value of the fine grained model was shown in [14] by analyzing of data from over 1000 students' two years usage of ASSISTment system. In [14], we presented evidence that, in general, the WPI-106 model did a better job at tracking students' knowledge and, thus, made a more accurate prediction of their end-of-year exam scores than the coarser grained models.

Table 1. Hierarchical relationship among transfer models.

WPI-106	**WPI-39**	**WPI-5**	**WPI-1**
Inequality-solving	Setting-up-and-solving-equations	Patterns, Relations, and Algebra	Math
Equation-solving			
Equation-concept			
X-Y-graph	Understand-line-slope-concept		
Congruence	Understand-and-applying-congruence-and-similarity	Geometry	
Similar-triangles			
Perimeter	Using-measurement-formulas-and-techniques	Measurement	
Area			

3.2 Approach

We have explained the nested hierarchical structure of our transfer models, and shown that the fine-grained model did the best *overall* at predicting student performance. Now we will examine our results more closely to see how different transfer models fit *different groups* of students.

3.2.1 Data

The dataset we use was collected during 2004-2005 school year. It involves 495 8th-grade students (approximately 13 years old) from two middle schools who have used the ASSISTment system on at least 6 days, with an average of 9 days. The item-level MCAS test report is available for all students so that we are able to evaluate accuracy of our models at state test score prediction. Since the scaffolding questions show up only if the students answer the original question incorrectly, students who answer the original question correctly do not have a chance at scaffolding questions, and would only be credited for the original question in the data. In order to avoid this selection effect, we preprocess the data using a compensation strategy to mark all scaffolding questions correct if a student gets an original question correct. Also, because our transfer models allow multi-mapping (one question associated with multiple skills), we choose to use a simple credit-blame strategy where if a student succeeds in answering a question, we mark all associated skills as being correctly applied, while when a student answers a question incorrectly, we only blame the weakest skill of the student, i.e. the skill on which the student has shown worst performance. After preprocessing, the data set contains 147,624 data points, among which 45,135 come from original questions. On average, each student answers 91 original questions. It is worth pointing out that during our modeling process, student response on original questions and scaffolding questions are used in an equal manner and they have the same weight in evolution.

The first portion of this research involves partitioning students into groups to determine if different groups of students have different patterns for learning math skills. Naturally, the 495 students can be separated by the schools they were in, with 312 from school F and 183 from school W. We also try to separate them by their performance level at the 2005 MCAS test. The high performing group includes the 128 students whose performance level is assessed by the state as “Advanced” or “Proficient”; the medium group includes the 154 students whose performance level is “Needs Improvement”, and the low performing group has the rest 213 students at performance level “Warning”. While these performance levels are somewhat specific to Massachusetts, they are at least criterion-referenced and much more general than numbers extracted from a student model or raw scores on a test (what qualifies as “Proficient” in Massachusetts is probably similar to “Proficient” in Macedonia). Our hypothesis is that students from a stronger school, or higher performing group, would show more transfer in their knowledge acquisition than those from a weaker school, or lower performing groups. Therefore, for the stronger students and schools the coarser grained model will better describe their learning and provide more accurate prediction of their MCAS test scores.

3.2.2 Modeling

In order to track individual student’s development of skills over time and make predictions, we choose to fit mixed-effects logistic regression models [8]. A mixed-effects model consists of both *fixed effects*, parameters corresponding to an entire population or repeatable levels of factors, and *random effects*, parameters corresponding to individual subject drawn randomly from a population. This approach takes into account the fact that responses of a student on multiple items are correlated. Moreover, the random effects allow the model to learn parameters for individual students separately. We use a logistic model because our dependent measure is dichotomous (0/1 for incorrect/correct). Regarding to the independent variables, for the fixed effects, we used a timing variable to represent the amount of time elapsed since the beginning of the school year, so that the model tracks the knowledge acquisition process longitudinally over time. Skills are included in the model as a factor to identify the skills associated with each response. Both the main effects of skills and an interaction term between the timing variable and skills are included in the model. Therefore, the model will learn an intercept (representing initial knowledge) and a slope (representing learning rate) for each skill separately. The timing variable is introduced as a random effect as well, in order to account for the learning rate variation of each individual student. The model is illustrated as below. To simplify the illustration, suppose TIME is the only covariate we care about in the model (*skill* can be introduced in a similar way). Thus, a 2-level representation of the model in terms of *logit* can be written as

Level-1 model: $$\log[\frac{p_{ij}}{1-p_{ij}}] = b_{0i} + b_{1i} * TIME_{ij}$$

Level-2 model: $$b_{0i} = \beta_0 + v_{0i}$$
$$b_{1i} = \beta_1 + v_{1i}$$

Where p_{ij} is the probability that student *i* gives a correct answer at the *j*th opportunity of answering a question;

TIME$_{ij}$ refers the *j*th opportunity when student *i* answered a question. In our data, it is a continuous value representing the number of months (assuming 30 days in a month) elapsed since the beginning of the school year.

b_{0i}, b_{1i} denote the two learning parameters for student *i*. b_{0i} represents the "intercept" or how good is the student's initial knowledge; b_{1i} represents the "slope" that describes the change (i.e., learning) rate of student *i*.

β_0, β_1 are the fixed-effects and represent the "intercept" and "slope" of the whole population average change trajectory.

v_{0i}, v_{1i} are the random effects and represent the student-specific variance from the population mean.

We fit the mixed-effects logistic regression models with R (http://www.r-project.org/) using the glmer() function in the *lme4* package [3], using "logit" as a link function. For simplicity, assuming knowledge was changing linearly (in logistic space) over time. One model is fit for each school and each performing group separately. Given a student's learning parameters on different skills, the skill-tagging of each MCAS question, and the exact test date of MCAS, we can calculate the probability of positive response from the student to each MCAS test question. Then we sum the probabilities up as the prediction of students' MCAS scores. Two prediction evaluating functions are chosen, mean absolute difference (MAD), and mean difference (MD), as below.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|MCAS_i - prediction_i\right| \qquad MD = \frac{1}{n}\sum_{i=1}^{n}(MCAS_i - prediction_i)$$

where *MCAS*$_i$ is the actual MCAS score of the ith student, and *prediction*$_i$ is the predicted score from our model. Both measures are used since MAD gives a good estimate the closeness of the prediction to actual scores while MD allows us to see if a certain model has been overestimating or underestimating.

3.3 Results and discussion

The results for both school F and school W are summarized in Table 2. As shown in Table 2, school F has a flat error line across all four different transfer models. The MAD for the WPI-39 model is the lowest, and yet a paired t-test that compares the absolute pair-wise differences of individual students among all models suggested that there is no reliable difference. However, for school W, the line tilts: the MAD of the WPI-39 model is reliably lower than those of the WPI-1 and WPI-5 models, indicating school W is better predicted by a finer grained model than by coarser grained models. Note that we are not able to fit the statistical model for school W with the WPI-106 transfer model (there is a technical glitch we do not understand and are investigating). We encounter the same problem later in the paper, which admittedly bring up some caveats in interpreting our results. The second part of Table 2 shows the values of MD for each model. The results indicate that both schools are optimized at the WPI-39 model. In general, student performance on the state test is overestimated by our models except that the WPI-106 model underestimates school F; and school W is even more overestimated than school F across known results from all the three models. As we know that, theoretically a one-skill model assumes perfect transfer. Since that is unlikely to happen, it would tend to overestimate student performance. And for a weaker school, perfect transfer is even more improbable. Thus, the overestimation would be greater since students are probably learning a collection of 106 unrelated skills. The tendency of overestimate decreases as the granularity of transfer models increases, and a very fine grained model such as the WPI-106 model that assumes no transfer or very low transfer may even underestimate when there is actually some level of knowledge transfer. We can see that in Table 2, the MD goes from negative to positive when we use the WPI-106 model for School F. Given these results, based on our hypothesis we would predict school F is the stronger school. An examination of both schools' MCAS performance reports (for current achievement) and information on their Annual Year Progress (AYP, for changes in performance) confirms our prediction.

Table 2. Results for students grouped by schools

Results	School	WPI-1	WPI-5	WPI-39	WPI-106
MAD	School F	4.188	4.168	4.124	4.175
	School W	4.669	4.601	4.329	N/A
MD	School F	1.362	0.932	0.477	-1.000
	School W	3.043	2.867	2.012	N/A

Table 3. Results for students grouped by performance levels

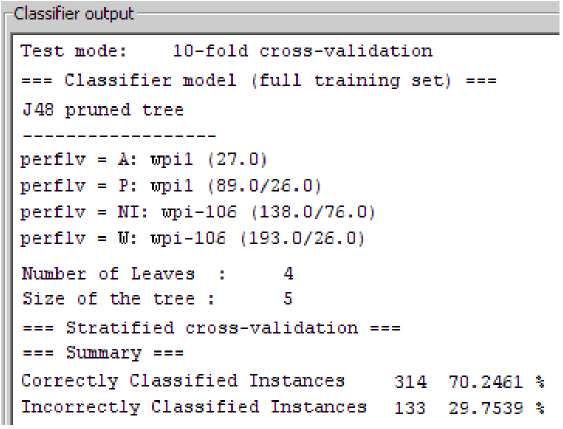
Results	Performance Level	WPI-1	WPI-5	WPI-39	WPI-106
MAD	Advanced/Proficient	2.673	2.834	2.489	3.249
	Needs improvement	3.180	3.243	2.900	N/A
	Warning	4.027	4.092	3.518	N/A
MD	Advanced/Proficient	-1.726	-2.034	-1.210	-2.715
	Needs improvement	1.534	1.744	0.893	N/A
	Warning	3.023	3.136	2.212	N/A

As mentioned in section 1, a second validation approach is that instead of partitioning students by school, we can use their state assessment test score and partition them by math proficiency. If we see a trend for stronger students, it is reasonable to believe it applies to stronger schools. Therefore, as reported in section 3, we split all the 495 students into 3 groups based on their state test performance level, and fit a mixed-effects logistic regression model to each group separately for different transfer models. The values of MAD and MD are summarized in Table 3. We see a slight support with MAD: for the students at the high end, the WPI-39 does the best job at predicting their state test scores, reliably better than the other three models, while the WPI-106 model does reliably worse than the WPI-1 and WPI-5 models, suggesting there is certain amount of knowledge transfer happening with the high performing students. However, since we do not obtain results of the WPI-106 model for the other two groups, it is hard to draw a conclusion there. When it comes to the MD measure, we notice some support as well. Obviously, the advanced and proficient students have been underestimated by all models, and the amount of underestimation goes worst when the finest grained model, the WPI-106 model, is applied. On the contrary, the medium and low performing students are all overestimated under all the models. Just as we hypothesize, the finer grained models overestimate less than the coarser grained models, and the better performing, stronger groups are less overestimated than the weaker groups. Therefore, weaker students are better represented by transfer models that are finer-grained.

3.4 A bottom-up aggregation approach

Rather than starting with an *a* priori disaggregation, we now focus on treating students as individuals and discovering commonalities among students who are best-fit with a particular transfer model. We have collected demographic data about several properties of a student, such as which school he/she goes to, ethnicity, gender, etc. Finding out the relation among these properties and which transfer model best fits this student is our goal. Our plan is to bring together model-fitting information and student characteristics, and then use a machine learning classifier to determine the best-fit model. This bottom-up aggregation is a strong alternative to proposing and testing disaggregation, and will scale nicely as we get more descriptors for each student.

For this purpose, we first re-fit models for all the students as one group[2] and identify which model best fits each individual student. The best-fit model information is then combined with other properties of the student in a new data set. Specifically, the properties we use are: gender, free-lunch status (indicative of family income), special education status, ethnicity, and state test performance level. These properties are picked because they are easy to access, and all of them have meanings to researchers working with other populations in other locations. In comparison, properties such as the school a student attends are much less useful to those in other locations. Given the new data set, we built a J48 (C4.5 revision 8) decision tree in Weka 3.6 [15]. The constructed J48 pruned tree is show in Figure 1 that tells how the classifier uses the attributes to make a decision. The constructed tree is extremely simple with just 5 nodes. The WPI-1 model is overall the best fitting model for Advanced (A) and Proficient (P) students, and the WPI-106 is for "Needs improvement" (NI) or Warning (W) level students. The numbers in brackets after the leaf nodes indicate the number of instances assigned to that node, followed by how many of those instances is incorrectly classified as a result. In our case, the correct classification rates are relatively good for students at performance level of A, P, and W. Yet, for students at performance level of NI, even though the WPI-106 model is the best fit, it is not dominant with 76 out of 138 instances misclassified. It is encouraging that this simple decision tree can achieve a predictive accuracy of over 70% during stratified cross-validation. Although the decision tree only uses MCAS performance, it was provided with the variables described above but was unable to find a use for them. This result suggests the appropriate level of transfer model granularity really seems to depend on student knowledge, rather than on variables that may correlate with knowledge such as family wealth. Therefore, if tutor designers have students with rather different levels of knowledge, they might wish to use different levels of their skill hierarchy. This point



```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
perflv = A: wpi1 (27.0)
perflv = P: wpi1 (89.0/26.0)
perflv = NI: wpi-106 (138.0/76.0)
perflv = W: wpi-106 (193.0/26.0)

Number of Leaves  :     4
Size of the tree :      5
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances     314    70.2461 %
Incorrectly Classified Instances   133    29.7539 %
```

Figure 1. Result of classifying in Weka

[2] We had to reduce the number of students to 447 from 495 because of a memory limit of R.

does not contradict the use of evaluating interventions [10] and schools by model granularity: other properties certainly matter in how well knowledge transfers, but for our dataset they are not as predictive as the student's knowledge.

4 Contributions, Future work, and Conclusions

The contribution of this paper lies in several aspects. First, automated techniques for revising transfer models for better knowledge representation have shown no huge improvements in accuracy but have addressed interesting scientific questions. Is there a way we can do interesting science on educational data sets and avoid the "irritating" automation step? Our answer is "yes," if it is possible to build a hierarchy of transfer models with different granularity. Previous experience tells us that this is not a rare thing to have, and not very hard to think about. The hierarchy can be used for runtime benefit of intelligent tutoring systems such as the control of mastery learning or generation of feedback messages for students of various proficiency levels. It can also be used to evaluate schools and be validated via high stake test performance. Second, through the usage of a bottom-up aggregation approach, the problem is changed. Rather than trying to automate the model search, why don't we automate seeing which student best fits which model? Third, we argue that hand-created transfer models and a bottom-up approach to aggregating students is a better use of human brains and computational power than approaches that focus search efforts on revising the domain model. Better understanding what parts of the scientific enterprise can be best done by people and which are better done computationally is a major issue in EDM.

A major open question of this work is whether just because a student is best modeled at a coarser grain size, shall we use such a model to drive tutorial instruction? For example, even though strong students are best modeled by a single skill "Math," it is not obvious how one would design hint messages in a system that only recognized one skill. A hybrid approach would be to track student knowledge and drive mastery learning at a coarser grain size, but provide feedback using a finer-grained model. A second question is that, since student knowledge is changing over time, perhaps we should use different level models to represent a student at different points in his learning?

In this paper, we start with existing hand-constructed transfer models at various levels of granularity, and use them as a lens to examine student learning. Specifically, we start by examining whether we can evaluate schools by determining the grain-size at which its students are best represented. We also examined what models best fit students at different levels of proficiency, and found some support for the idea of stronger students being better fit with coarser transfer models. The most interesting analysis was the bottom-up aggregation and using classification to find clusters of students who learn similarly. This analysis suggests transfer model granularity really seems to be about student knowledge. Finally, we argue that it is more productive to focus analytical effort on which students should use which transfer models rather than on automatically refining those models.

Acknowledgements

We thank Dr. Neil Heffernan for his help and insightful comments on this work. This research was made possible by the U.S. Department of Education, Institute of Education Science (IES) grants #R305K03140 and #R305A070440, the Office of Naval Research grant # N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All the opinions, findings, and conclusions expressed in this article are those of the authors, and do not reflect the views of any of the funders.

References

[1] 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.*

[2] Barnes, T. (2006). Evaluation of the q-matrix method in understanding student logic proofs. *Proceedings of the 19th International Conference of the Florida Artificial Intelligence Research Society* (FLAIRS 2006), Melbourne Beach, FL, May 11-13, 2006.

[3] Bates, D. (2007). Linear mixed model implementation in *lme4*. University of Wisconsin, 15 May 2007.

[4] Birenbaum, M., Kelly, A., & Tatsuoka, K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education*, 24(5), 442-459.

[5] Cen, H., Koedinger, K, & 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. Menlo Park, California: AAAI Press. pp. 47-53.

[6] Cen, H., Koedinger, K., & Junker, B. (2006). Learning factor analysis – A general method for cognitive model evaluation and improvement. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems.* Springer-Verlag: Berlin. pp. 164-175.

[7] 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.

[8] Hedeker, D. & Gibbons, R. D. (2006). Longitudinal Data Analysis. Hoboken, NJ: John Wiley & Sons.

[9] Koedinger, K. & Junker, B. (1999). Learning factor analysis: Mining student-tutor interactions to optimize instruction. Presented at Social Science Data Infrastructure Conference. New York University. November, 12-13, 1999.

[10] Koedinger, K. R., & Mathan, S. (2004). Distinguishing qualitatively different kinds of learning using log files and learning curves. In the Working Notes of the ITS2004 Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes.

[11] Leszczenski, J. M., & Beck, J. E. (2007, July 9). What's in a word? Extending learning factors analysis to modeling reading transfer. *Proceedings of the AIED2007 Workshop on Educational Data Mining*, Marina del Rey, CA, 31-39.

[12] Newell, A, & Rosenbloom, P.S. (1993). Mechanisms of skill acquisition and the law of practice. In P. S. Rosenbloom, J. E. Laird, & A. Newell (Eds.), *The Soar Papers: Research on integrated intelligence*. Cambridge, MA: MIT Press.

[13] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, pp. 555-562. Amsterdam: ISO Press.

[14] Feng, M, Heffernan, N., Heffernan, C. & Mani, M. (in press). Using mixed-effects modeling to analyze different grain-sized skill models. To appear in *the IEEE Transactions on Learning Technologies Special Issue on Real-World Applications of Intelligent Tutoring Systems.*

[15] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/

[16] Winters, T., Shelton, C., Payne, T., & Mei, G. (2005). Topic Extraction from Item-Level Grades. In Beck. J. (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Menlo Park, California: AAAI Press. pp. 7-14.