Pavlik Jr., P. I., Cen, H., & Koedinger, K. R. (2009). Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In T. Barnes, M. Desmarais, C. Romero & S. Ventura (Eds.), Proceedings of the The 2nd International Conference on Educational Data Mining (pp. 121-130). Cordoba Spain.

Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models

Philip I. Pavlik Jr.[1], Hao Cen[2], and Kenneth R. Koedinger[1]
{ppavlik, hcen}@andrew.cmu.edu, and koediger@cmu.edu
[1]Human Computer Interaction Institute, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University

Abstract. This paper describes a novel method to create a quantitative model of an educational content domain of related practice item-types using learning curves. By using a pairwise test to search for the relationships between learning curves for these item-types, we show how the test results in a set of pairwise transfer relationships that can be expressed in a Q-matrix domain model. Creating these Q-matrices for various test criteria we show that the new domain model results in consistently better learning curve fits as shown by cross-validation. Further, the Q-matrices produced can be used by educators or curriculum designers to gain a richer, more integrated perspective on concepts in the domain. The model may also have implications for tracing student knowledge more effectively to sequence practice in tutoring/training software.

1 Introduction

Because of the complexities involved in curriculum design, and because of the possibility of expert blind spots in the design of curricula [1], an alternative to human sequenced curricula might be desirable. One way to achieve this goal is to create a model that explicitly captures the pairwise knowledge component (KC, which may refer to skills, procedures, concepts or facts) relationships between item-types in the domain, what might be called a "transfer model". While a model that captures transfer may have many forms, here the transfer model we will investigate maps to a Q-matrix, which is a matrix in which rows represent item-types and columns represent KCs. In such a matrix, a 1 value indicates the item-type uses the KC, while a 0 indicates the KC is not involved in the item-type performance. Such a model is desirable because it allows us to make determinations about the optimal order of problems (sequence of repetition and presentation) since it allows prediction of which item-type will cause learning of KCs that transfer to the other item-types most efficiently. (Throughout this paper we use the term item-type to signify a collection of practice items that are either identical or only slightly different, e.g. instantiated with different numbers of the same relative magnitude but involving the same numerical operations or concepts.)

Many others have worked on such models of domains using testing data results [e.g. 2, 3]; however, this paper attempts to broaden the problem by simultaneously addressing the issue of what learning data results imply for the domain model found. It seems that a solution to this problem would have to propose some sort of "transfer function" that captures how the learning and performance of item-type A causes effects on the learning and performance of item-type B. Indeed, learning transfer function models already exist (logistic regression models that predict item-type B as a function of prior practice of either A and/or B) and have been used with human generated KC sets for the purpose of refining the model of individual KCs [4, 5]. A similar issue regarding domain structure

and learning curves has been addressed by others who have looked at how to use human derived domain structures to combine learning curves to modify adaptive instruction so that feedback is based on combined KCs rather than the individual KCs [6, 7].

By combining these two approaches (machine derived domain modeling with learning curve analysis of transfer) we hope to produce a technique that both avoids the need for possibly error-prone, time-consuming human labeling of the domain and avoids the problems inherent in proposing an automatic domain model using observations that are non-static (i.e. they change with time as a function of learning). For this paper we will constrain our focus to explaining our pairwise-method of transfer testing, and how this can be used to automatically determine an overall linked (transfer) model of a domain using learning curves. We will then compare this linked model (with various criteria for linkage) with an unlinked model. Cross-validation will be used to establish the superiority of the linked model at various linkage criteria.

1.1 POKS and LiFT (Learning Factors Transfer)

This work is conceptually similar to work with knowledge spaces using the partial order knowledge structure (POKS) method [8]. In both cases we construct a partial order directed acyclic graph using pairwise tests to determine linkages and their directionality. In POKS the relationship between 2 item-types (A and B) is written A→B, and allows inferences of the type "if A is known, then B must be known" and "if B is unknown, then A is unknown". In contrast, in the Learning Factors Transfer (LiFT) test a directional relationship is expressed in set notation so an analogous link is written $A \supset B$ and expresses the inference that the KCs required for A are a proper superset of the KCs required for B. For instance, in the dataset we examine, item-type A might be “What is 1/1 in percent? (Answer: 100%)”, which requires an understanding of whole number fractions and the percent conversion procedure, whereas item-type B is “What is 1/1 in decimal? (Answer: 1)” and only requires the understanding of whole number fractions. Because this example represents the superset relationship, it encodes a situation where learning of item-type B transfers fully to A, but where learning of item-type A transfers only partially to B. This example describes a model with 2 KCs for A and 1 KC for B. Thus, practice of B benefits A partially, while practice of A benefits both A and B.

The POKS test is not fully applicable to repeated practice opportunities for a specific item-type because it requires single observations for each item from each subject to compute in standard form (the test assumes that observations are independent). In previous work we showed how it was possible to use the average performance rather than a single observation to compute implication relationships using POKS [9]. While this POKS analysis provided interesting information about the domain, like many previous works describing domain structure, the result necessarily abstracted over learning effects which the LiFT test explicitly analyzes. This inclusion of the effect of learning is a key advantage of our new method, since the LiFT test (assuming it shows transfer between 2 item-types) should therefore allow us to better answer questions such as which item-type should be practiced first and how much it should be practiced before it is optimal to switch to the other. For example, in some cases the quantitative model (Section 2.1) will predict, given $A \supset B$, that B should be practiced because initial performance of A will be

poor without some practice of B first. In other cases, A might be easy enough or the prior learning of B might be strong enough that A should be practiced first. Additionally, it seems plausible to suggest that if our test function does not include a learning component then a domain search using it will be less accurate when our data contain significant amounts of learning.

2 Learning Factors Transfer (LiFT) Analysis

LiFT analysis takes the form of a basic pairwise test that establishes the likelihood that any pairwise relationship is better represented as a transfer relationship, or whether it appears the item-types are unrelated. While the results of the pairwise test may be useful for human curriculum designers to consider, the LiFT test can also be employed to mine large datasets with many item-types being learned simultaneously. We will describe how this algorithmic usage can be performed, and compare the model fit for the transfer relationships discovered at various criterion settings with 3 alternative models: a single item/KC model, a independent item/KC model, and a random transfer models with the number of links in the Q' matrix yoked to the LiFT found transfer models, but placed at random. Because transfer is a within-subject effect, occasionally we will see it occur due to general correlation among item-types caused by latent variables such as motivation, general intelligence, or generally better prior learning. This possibility may be minimized by setting a strict criterion for our transfer test.

2.1 PFA Item-type model

The model equation we will use here is similar to a model equation that has been recently described and shown to fit better than either the standard logistic regression equation used in LFA (Learning Factor Analysis) or the standard version of Bayesian knowledge tracing [10]. For this paper we have made a slight modification which specifies that prior learning parameters are assigned at the item-type level rather than the KC level. The model equation, which can be referred to as the Item-type PFA equation, is a logistic regression model of performance on each trial that includes a parameter to capture the initial strength for each item-type and 2 parameters that track the learning and performance with each KC. The PFA Item-type equation is shown in Equation 1, where *m* is a logit value representing the accumulated learning for a student *i* on one item-type *k* using one or more KCs *j*. The easiness of the item-types is captured by the β parameter for each item-type. The effect of learning and performance is captured by *s*, which tracks the prior successes for the KC for the student and *f*, which tracks the prior failures for the KC for the student. The 2 parameters γ and ρ scale the effect of these observation counts for each KC as a function of the *s* or *f* of prior observations for student *i* with KC *j*. Equation 2 is the logistic function used to convert *m* strength values to predictions of observed probability. It is useful to note that the model always assumes that each item-type has a "base KC" that matches to each item-type and the LiFT test tries to go beyond that base KC to propose other KCs that might be transferred to the item-type to better model performance.

$$m(i, j \in KCs, k \in Items, s, f) = \beta_k + \sum_{j \in KCs} (\gamma_j s_{i,j} + \rho_j f_{i,j}) \quad (1)$$

$$p(m) = \frac{1}{1 + e^{-m}} \quad (2)$$

The assignment of KC's to item-types is described by a Q-matrix which describes which KC's influence which item-types. To represent an independent component for each item-type we have specified that each KC is matched to a specific item-type. Therefore, our Q-matrix will be a square matrix with item-types as rows and matched KCs as columns. Since every item-type has its matching KC, the diagonal will be all 1s, indicating that each KC is present in its corresponding item-type. To distinguish this kind of Q-matrix (square where every item-type is assigned at least 1 KC) from the larger set of standard Q-matrices, henceforth we will refer to it as the Q'-matrix.

2.2 LiFT test

The LiFT test takes as input the sequence of practice data for 2 item-types in a tutor (their learning curves) and computes the relative likelihood that they have an overlapping KC by comparing the weights of the likelihood difference of alternative models. In the version we are presenting here, we considered the 2 alternative models, A ⊃ B and A~B (where A and B do not share a KC) for each order pair of item-types (thus pair (X, Y) is distinguished from (Y, X) and the test is run on both pairs). While we consider these 2 models, there are other transfer assumptions that might be tested but these are beyond the scope of this paper. The A ⊃ B model asserts that the A item-type is controlled by the 2 KCs, while the B item-type is only controlled by 1 KC (the 2x2 Q'-matrix is filled with 1's on the diagonal and the upper right corner is also filled with a 1 to indicate item-type A shares the same KC as item-type B). In contrast, the A~B model supposes that each item-type is independent (the Q'-matrix is only filled with 1's on the diagonal since each item-type has a single KC).

When this test is computed it determines whether we can get an improvement in model fit by proposing that learning for one item-type transfers to the performance of the other item-type. At the same time as it determines whether the item-types share a KC, the directionality of the test determines which of the two item-types contains an additional independent KC, thus indicating that it contains a superset of the skills required relative to the other item-type. To compute the test, we fit these 2 models (each with 6 parameters) and compared them according to their BIC (Bayesian Information Criterion) weights to determine the evidence ratio in favor of A ⊃ B. (Because model complexity was equal, this was equivalent to using AIC weights or likelihood ratio.) This evidence ratio gives us the likelihood of A ⊃ B expressed as a probability. This use of BIC weights to compute evidence ratios has been described in detail previously [11]. Because the BIC weight test requires observations to be independent, we minimized BIC using the average loglikelihood for each subject rather than for each observation. This procedure is conservative since it overcompensates for the only partial dependence between observations within a single subject.

2.3 LiFT algorithm

Many possibilities exist for how this test could be applied to a dataset to determine the relationships between item-types. For this first attempt we did an exhaustive search for pairwise relationships, accepting those BIC weight test results that resulted in improvements above a probability criterion. Acceptance of the result of any pairwise test meant that the superset transfer implication was added to the Q'-matrix. For example, imagine that the criterion is .43 and we test $A \supset B$ and get a probability value of .32. In this case we have not passed criterion and we do not alter the Q'-matrix row for item-type A. However, if the $A \supset B$ test arrived at a result of .87 (thus passing the criterion), we would enter a 1 in the Q'-matrix at item-type row A and KC column B.

This LiFT test is applied to a multi-item-type dataset according to the following steps.

1. Create diagonal matrix with 1's on the diagonal assuming rows and columns equal the number of item-types.

2. Compute the pairwise test for all non-diagonal entries, entering a 1 in the matrix for any tests that pass.

3. Use the Q'-matrix from step 2 and maximize the likelihood of the entire dataset.

Because we wanted to get a perspective on what was an effective criterion, we tested criteria from 0 to 1 in .01 increments. We used 10 fold cross validation of the mean absolute deviation averaged by subject to compare the full models we tested. In applying this algorithm we used the following dataset.

3 Data

The dataset we used was gathered from a middle school in Florida which uses the Bridge to Algebra Cognitive Tutor by Carnegie Learning Inc. As part of a larger investigation we had supplemented the tutor with 9 problems sets each with 34 item-types. These supplemental units were closely matched to the tutor units that followed them, and future analysis will look at the potential for transfer into the Bridge to Algebra tutor from the supplemental units. For this investigation we choose 1 of these supplemental units (Unit 5, Fraction, Decimal and Percent Conversions) to investigate how our algorithm would work to improve the predictions of the model by filling in the Q'-matrix. The item-types were ideally sequenced for an analysis of this sort since they were randomized into a 4 by 2 within subjects design where there were 4 levels of practice (0, 1, 2 or 4 repetitions) and 2 levels of spacing (3 or 15 intervening trials). These conditions (with a few buffer trials) resulted in total of 64 practices for each supplemental lesson. 361 students produced valid data. The data for each subject took the form of sequential lists of which item-type was practiced and whether the result for that practice was correct or incorrect.

4 Results

Figure 1 shows the size of the resulting Q'-matrices found by the LiFT algorithm run on Florida dataset. At a criterion of 0, every BIC weight test passes and the matrix is saturated with 1 values. In this case the algorithm fits a model with 34 item-type parameters (β) and 34 γ and ρ parameters which are identical for every KC because they are shared across every item-type (a square Q' matrix filled with 1s). As the criterion is made more stringent, Figure 1 shows how fewer and fewer links are proposed until at a criterion of 1, none of the BIC weight tests pass, and the algorithm proposes 34 item-types, each with an associated KC, each represented by its own β, γ, and ρ (a diagonal Q' matrix of 1s).

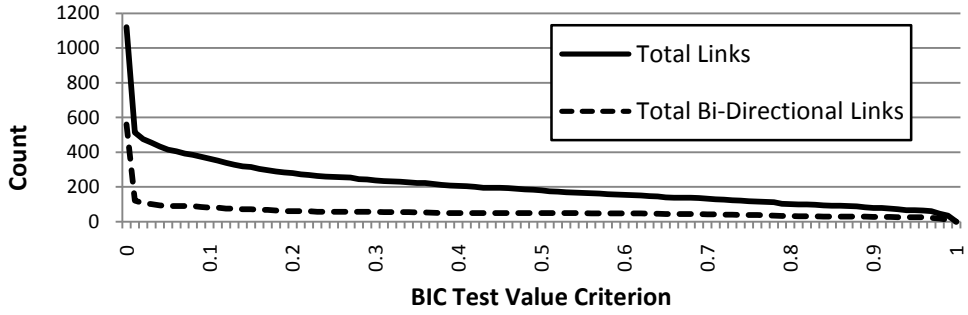

Figure 1. Number of total links and number of bidirectional links A ⊃ B found with pairwise BIC weight test.

Figure 2 shows 10 fold cross validated estimates of mean absolute probability deviation (averaged by subject) for the model fit at each criterion. For comparison, the 3 alternative models are also presented on this figure. The 1 KC/item model comparison is a 3 parameter model which assumes that all of the 64 practices for each student are actually best modeled as a single item-type with one KC. The no transfer model comparison is a 102 parameter model with 34 KCs/items and is represented by a diagonal Q'-matrix which assigns 1 KC to each item-type (equivalent to criterion = 1). The yoked random control comparison is a model calculated with a random square Q'-matrix yoked to the number of links in the LiFT found Q'-matrix at that criterion, but with those links placed randomly. Note: Just as with the found Q'-matrices, we began with the assumption that each item-type was associated with at least a single KC. The yoked random control is an important comparison since it helps to establish that selection using the LiFT test is causing the advantage seen, rather than it being caused merely by the presence of links in the Q'-matrix.

The result shown in Figure 2 establishes that the transfer model (LiFT found Q') fits the data better than the 1 KC/item, no transfer, or yoked random Q'-matrix models. Further, the cross validated comparison establishes that the result is likely to generalize to similar populations. Of some interest for further research is why the improvement in fit is relatively large even when using an extremely liberal BIC weight test value criterion. It seems likely that averaging of multiple low criterion transfer relations (multiple 1s in a Q'-matrix row) causes a reduction of the error compared to the pairwise test models.

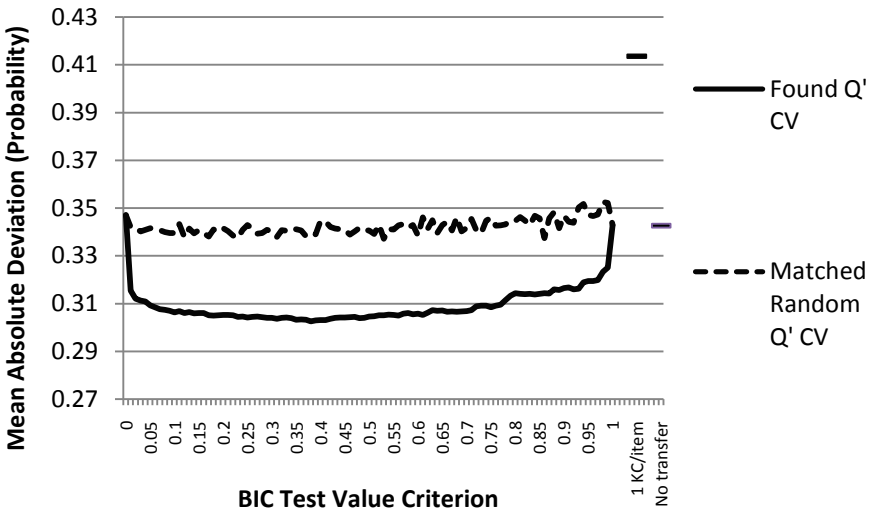

Figure 2. 10-fold cross-validated fit of the model and yoked random control across the range of BIC weight test criterion values.

4.1 Interpretation of Results

In explaining the LiFT test, we provided an example where the two item-types showed an hypothetical subset relationship with one item-type requiring 1 KC and the other requiring that KC plus an additional KC. However, we did not find any clear superset relationships when we looked at the results. In contrast to this theory, our results were more varied but showed several specific patterns. We examined these patterns at a few test value criteria finding quantitative but not qualitative differences. The following description is given for a criterion of 0.6 on the BIC weight test.

Group 1 (4 item-types) was the smallest item-type group where item-types were left completely unlinked. Because these item-types were unlinked, the found Q'-matrix models of these item-types (determined by β, γ, and ρ and Equation 1) is identical to the no transfer model of these item-types. This set includes questions that are relatively difficult for most students (a low β parameter). However, these item-types also tend to have high learning and performance parameters (a high γ and ρ). Based on these results we might suppose that the problems are badly worded, tricky, or beyond the level of the average student. For instance, the item-type (instantiated with different numerals in 6 versions that were randomly selected from with replacement) "If we are given that 4% of y is 1, one fraction of our proportion is 4/100 and the other is what? (Answer: 1/y)" was found to be in this category. Compared to other item-types it is wordy and highly complex involving fraction KCs, percent KCs, and proportion solving KCs.

Group 2 (15 item-types) was similar to Group 1 in that these item-types did not share their KC component with any other item-types. Unlike Group 1 however, these item-types had between 2 and 10 input KC's that other item-types shared with them. As we will see, Group 3 provides these inputs KCs. Group 2 item-types are distinguished by being primarily repetition based vocabulary practice item-types. Further, we see less

fundamental conversions ("What is 1/1000 in percent? (Answer: 0.1%)" and "What is 10/1 in percent? (Answer: 1000%)") are included in this group, perhaps because their performance and ability to be learned depends on understanding more fundamental frequency conversions (e.g. 1/10 in decimal), while the converse is less true.

Group 3 (15 item-types) was composed of item-types that were strongly connected with other item-types both by sharing their matched KC with between 2 and 22 item-types and by receiving KC input from between 3 and 11 item-types. Group 3 is composed of item-types that start out at moderately higher βs (logit greater than 0, i.e. >50% initial performance) and have much lower γ and ρ parameters for their matched KCs. These item-types appear to describe 3 conceptual categories that can be distinguished by the dominant sharing of inputs and outputs within each category. Essentially what has happened in each case is that inputs and outputs form loose conceptual categories by sharing that is primarily within category. Category 3a includes two item-types "What is a ratio of the amount of decrease to the original value, written as a percent? (Answer: percent decrease)" and the corresponding question about an increase. This category seems to depend on the general form of the question (identical) which allows direct transfer of the solution pattern. Category 3b is similar, but has to do with 3 questions that required the partial solution of proportions (e.g. "Given the proportion 1/y = 3/4, what is the product of the extremes? (Answer: 4)") One of these questions also interacted heavily with category 3c. Category 3c can most closely be aligned with a fundamental proficiency in the domain, or with general intelligence, since these item-types all shared their KC with at least 13 item-types and received inputs from at least 8 item-types. While each of these item-types had its own β parameter, this extensive sharing means that learning and performance change for these item-types occurs nearly as a unit. 8 of these item-types involved simple place value problems such as "What digit is in the 10s place in the number 25046.37189? (Answer: 4)" and "What is 1/10 in decimal? (Answer: 0.1)" Also in this category was a simple pre-algebra problem: "What is y in the equation $3 \times y = 9$? (Answer: 3) (with 6 versions). Finally, the item-type "If 54 can be completed in 10 hours, what is the amount completed in 1 hour (as a decimal)? (Answer: 5.4)" (with 6 versions) appeared to be part of both 3b and 3c, which seems consistent, since it involves aspects of both proportions and place value (the set of item-types always required a power of ten calculation).

The three groups found show that understanding the model has interesting implications for the curriculum designer. First, it appears that item-types in Group 1 need to be improved in some way. While we cannot tell for sure what the problem is, the lack of connections to the other groups establishes that these item-types are either in a different domain, too complex to relate to the more simple item-types in the set, or badly worded so that students cannot transfer in related knowledge. Group 2, items on the other hand receive KC from Group 3 item-types, so we can surmise that they are at least marginally related to the domain. The fact that these questions did not share their KCs might have more to do with the limitations of the question set overall (which did not require much application of these mostly vocabulary-based item-types in other questions) or the model (see Discussion) rather than any specific lack in the items themselves. Finally, Group 3 items are useful to consider because the model here suggests that the 3 conceptual categories (3a, 3b and 3c) are each better modeled as single units rather than as

independent skills. Knowing these item-type clusters are closely related allows curriculum designers to have more information as they make curriculum design decisions.

5 Discussion

In general, the results were a qualified success because the model found produced a better fit that generalized and because the model structure provided other reflections on the content subdivisions in the domain of item-types studied. Despite this, there were problems with the current logistic regression model. Specifically, because of the way the equation uses the Q'-matrix to share KCs as whole units, the parameter magnitudes changed as a function of whether or not each item-type shared KCs with other item-types. Item-types that shared their KC's (especially when they provided their KC to many other item-types) had lower performance parameter values (γ and ρ) than when they were fit in the no transfer model. The reason for this is simply that when the probability is determined from multiple inputs each input must be scaled down so the total growth still resembles the situation with one KC. However, one unfortunate consequence of this is that the model is then insensitive to repetition effects for a single item-type and therefore predicts much slower growth when the same item-type is repeated. This is because the item-type is no longer being controlled by a single KC, but rather has become tied to a collection of KCs that control it. One solution to this problem in future work may be to take a knowledge decomposition approach that does not insist on this KC sharing through the Q'-matrix [12]. Such an approach might instead propose that the magnitude of transfer for each KC is different depending on what item-type the KC transfers to. While this would add parameters to our model, it might also greatly improve the fit of the model.

This work may apply directly to the educational problem of sequencing item-types to maximize learning because the resulting model captures learning, is adaptive to performance, and captures the domain structure together in a single model. Unlike other automatically determined domain models, which determine performance dependencies and might be used for ordering practice, our model explicitly tracks learning also. By adding learning to our domain model, our model has the potential to answer not just the question of what item-type is best next, but also the question of how much more should the current item-type be practiced.

Acknowledgements

This research was supported by the U.S. Department of Education (IES-NCSER) #R305B070487 and was also made possible with the assistance and funding of Carnegie Learning Inc., the Pittsburgh Science of Learning Center and DataShop team (NSF-SBE) #0354420, and Ronald Zdrojkowski.

References

[1] Koedinger, K., Nathan, M.J.: The real story behind story problems: Effects of representation on quantitative reasoning. *Journal of the Learning Sciences*, 2004, 13(2)**,** p. 129-164.

[2] Barnes, T.: The Q-matrix Method: Mining Student Response Data for Knowledge. *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, 2005.

[3] Spacco, J., Winters, T., Payne, T.: Inferring use cases from unit testing. *AAAI Workshop on Educational Data Mining*, 2006. ACM Press.

[4] Leszczenski, J.M., Beck, J.E.: What's in a Word? Extending Learning Factors Analysis to Model Reading Transfer. *13th International Conference on Artificial Intelligence in Education, Educational Data Mining Workshop*, 2007. Los Angeles, CA.

[5] Cen, H., Koedinger, K.R., Junker, B.: Learning Factors Analysis - A general method for cognitive model evaluation and improvement. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 2006. Springer Berlin / Heidelberg, p. 164-175.

[6] Martin, B., Mitrovic, A.: The effect of adapting feedback generality in ITS. In: Wade, V., Ashman, H., Smyth, B. (Eds.) *AH 2006*, 2006. p. 192-202.

[7] Martin, B., Mitrovic, A.: Using learning curves to mine student models. *10th International Conference on User Modelling*, 2005. Edinburgh, p. 79-88.

[8] Desmarais, M.C., Maluf, A., Liu, J.: User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction*, 1996, 5(3-4)**,** p. 283-315.

[9] Pavlik Jr., P.I., Cen, H., Wu, L., Koedinger, K.R.: Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor. In: Baker, R.S., Beck, J.E. (Eds.) *Proceedings of the 1st International Conference on Educational Data Mining*, 2008. Montreal, Canada, p. 77-86.

[10] Pavlik Jr., P.I., Cen, H., Koedinger, K.R.: Performance Factors Analysis -- A New Alternative to Knowledge Tracing. In: Dimitrova, V., Mizoguchi, R. (Eds.) *The 14th International Conference on Artificial Intelligence in Education*, 2009, accepted. Brighton, England.

[11] Wagenmakers, E.-J., Farrell, S.: AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 2004, 11(1)**,** p. 192-196.

[12] Zhang, X., Mostow, J., Beck, J.E.: All in the (word) family: Using learning decomposition to estimate transfer between skills in a Reading Tutor that listens. *AIED2007 Workshop on Educational Data Mining*, 2007. Marina del Rey, CA.