What's in a Word? Extending Learning Factors Analysis to Model Reading Transfer

James M. LESZCZENSKI and Joseph E. BECK
School of Computer Science
Carnegie Mellon University
Project LISTEN, Newell Simon Hall, 5000 Forbes Avenue Pittsburgh, PA 15213

Abstract. Learning Factors Analysis (LFA) has been proposed as a generic solution to evaluate and compare cognitive models of learning [1]. By performing a heuristic search over a space of statistical models, the researcher may evaluate different cognitive representations of a set of skills. We introduce a scalable application of this framework in the context of transfer in reading and demonstrate it upon Reading Tutor data. Using an assumption of a word-level model of learning as a baseline, we apply LFA to determine whether a representation with fewer word independencies will produce a better fit for student learning data. Specifically, we show that representing some groups of words as their common root leads to a better fitting model of student knowledge, indicating that this representation offers more information than merely viewing words as independent, atomic skills. In addition, we demonstrate an approximation to LFA which allows it to scale tractably to large datasets. We find that using a word root-based model of learning leads to an improved model fit, suggesting students make use of this information in their representation of words. Additionally, we present evidence based on both model fit and learning rate relationships that low proficiency students tend to exhibit a lesser degree of transfer through the word root representation than higher proficiency students.

1. Introduction

The task of modelling student cognition is at best an attempt to approximate how students mentally represent a given task. On one hand, the researcher can never know exactly how a student internally represents a problem, since often the student might not be exactly aware either. Approaches such as think aloud protocols have been proposed to capture such representations, but are generally too expensive and time-consuming to use in anything but limited studies [2]. However, a reasonably simple assumption can be made in this respect: if we build two models and one of them is a better fit to natural student learning data, we can say that it is a closer approximation to the student's mental representation. The challenge lies in specifying a good model. Even for simple, well-defined tasks, the obvious representation may not always be the right one. Symbolization in word problems, for example, has led to some counter-intuitive results concerning how easy certain mental representations are for students [3]. In fact, prior research has provided evidence that the choice of a model and its basic knowledge components is nontrivial. Determining the best choice of meaningful knowledge components, however, remains beyond the scope of this paper [4].

Learning Factors Analysis (LFA) has been proposed as a generic solution to evaluate and compare many potential cognitive models of learning [1]. As input, LFA takes an initial model of student representation and a set of legal operators to transition from one model to another. By performing a heuristic search over statistical models, the researcher may evaluate different cognitive representations of a set of skills. This framework offers the capability to compare different representations quantitatively, in terms of the statistical model chosen. Through such manipulation, we can check many hypotheses about what skills have the most in common. The advantage to this approach is that it bases all decisions about which model is best on just the data and the transition operators. As was shown in [1], LFA can locate skills that contain such commonalities.

From the perspective of a student's representation, the research task is to approximate how transfer occurs. Given an accurate model of a student's representation, such transfer would be easily extrapolated. To do so, the

commonalities between skills (and thus the degree of transfer) would be implicit. The LFA framework attempts to model these relationships between skills under the assumption that if two skills A and B are better modeled by a single skill, then practice on either A or B transfers to the other. This implies that transfer is symmetric within an equivalence class of similar skills. The goal of the search is to identify these equivalence classes.

Above all, LFA has a clear advantage in that the underlying approach is domain-independent. The statistical model has no special knowledge about any particular field; it merely requires that the domain is defined by some set of skills, or knowledge components, which are then manipulated strictly based upon the empirical data. Thus, any advances in the performance, heuristics, or even the statistical model itself can be directly applied to all such domains.

One domain with perennial significance in the field of education is reading, where knowledge transfer has a clearly defined meaning. If a student reads a word, are there other words in the language at which we believe he may have become more proficient? One extreme approach would be the "bag of words" model, where every word is entirely independent. Reading a word in this representation will have no bearing on any other. However, this approach seems lacking, or else teaching students to read would simply consist of memorization of a significant subset of the English language, word by word. At the other extreme, one might suggest that every pair of words has some dependence. This implies that every time a student reads a word, he becomes (marginally) better at every other word in the language. This alternative approach also seems counterintuitive, as the authors feel that a student, both before and after reading the words 'dog' and 'cat' for the first time, will fail equally quickly on a word like 'supercalifragilisticexpialidocious'. However, we could reasonably assume that the student would perform more accurately on 'dog**s**' and 'cat**s**', relative to a similar student with no practice on any of these words.

In this paper, we describe a scalable application of the LFA framework in the context of knowledge transfer in reading. In this paper, we refer to 'words' and 'skills' interchangeably, as words are the initial building blocks of our cognitive models. Our intent is to develop a model that can identify students' representations of reading from data, as opposed to a theory-based assumption that students are all the same. We scale an existing educational data mining method to a level feasible for extensive reading data, and observe the resulting representations.

Given this architecture for reading analysis, what's in a word? Reading is certainly a task, and a task can theoretically be separated into knowledge components. Identifying these components is the complication, which an LFA-based approach can decide based solely on the data. This paper explores how this framework can extend to a corpus of reading data and makes observations about the implications of the resulting models with respect to students' mental representations.

2. Application to Reading Tutor

In order to provide a corpus of reading data for the LFA framework, we selected data from the Project LISTEN Reading Tutor in the 2003-2004 school year. The Reading Tutor [5] is a type of intelligent tutoring system (ITS), which assists elementary school kids in learning how to read. Using speech recognition technology, the Tutor listens to and records student utterances with which it compiles a database of information relating to student performance [6]. Additionally, it attempts to help students based on its recognition of these utterances.

Users of the Tutor were elementary school students in grades 1-6, mainly from the Pittsburgh, PA area. Paper tests were given both at the beginning and end of the year to assess individual reading level and improvement over the course of the school year. In 2003-2004 alone, the Reading Tutor collected approximately 6.9 million attempts by 650 elementary school students to read words. It is worth noting that the correctness of an attempt is defined as whether the automated speech recognizer (ASR) decided whether the student spoke the correct word. Analyses show that ASR correctly flags around one quarter of misread words, with a false positive rate of around 4% [7]. Among the additional data collected by the tutor are a timestamp to maintain a temporal ordering among a student's interactions with the Tutor, information about help asked for by the student, and latencies between utterances of words.

In an attempt to minimize the noise in our results, we chose to pre-process the database and screen out instances that seemed to poorly represent the trends present in the data. To this end, we first chose to only consider instances where a word was being read by a student for the first time in a day. This decision is supported by the observations in [8], where it was noted that "in general, massed practice is not helpful to learning." Additionally, we chose to screen out a list of 36 common stop words to avoid placing a great deal of weight on words that were already mastered by most students, and for which it would likely be difficult to discern a learning curve. To reduce the dimensionality of the data and avoid problems with sparsely estimated parameters, we screened out all attempts where there were not at least five students who had encountered the word at least five times (subject to the above constraints). Finally, since little gain in word decoding skill is to be expected after the word is read many times, we only considered the first 50 exposures

to each word by a student. This screening process resulted in a set of 651,301 attempts by 469 students to read 1011 distinct words.

The advantages to using the Reading Tutor as a data source are numerous. Above all it is ecologically valid, in the sense that all data collected are from a natural learning environment, unhindered by confounding effects that might be present in a laboratory setting. The autonomous nature of data collection avoids both grader bias and inconsistency. Additionally, it allows for longitudinal, fine-grained, and comprehensive data to be collected with minimal effort, as opposed to a more typical and controlled psychological study where data generally must be aggregated by hand. The obvious trade-off is the inaccuracy associated with ASR, but this is quickly outweighed by the ease with which 6.9 million natural data points can be collected.

3. Background Theory and General Approach

Learning Factors Analysis (LFA) has been proposed as an interface by which to "combine statistics, human expertise and combinatorial search to evaluate and improve a cognitive model" [1]. This framework represents students' learning in terms of a logistic model, and performs a heuristic search over the space of all such models to locate the best of them. A discussion of model scoring heuristics will be saved for later in the section. The models themselves are each based upon a set of data including the students, skills performed, and the outcome of each trial. Models are transformed via 'split', 'add', and 'merge' operators which act on a single skill at a time, making incremental transitions between similar models. In essence, the operators are re-labelling the instances of data to form a new dataset, derived from the original. Over these potential states we perform a best-first search, where we further explore the states which seem to be the most beneficial to us. It is worth noting that while we direct our search in this fashion, countless other methods could be chosen to direct the heuristic search, such as the association rules described in [9].

Before describing a representation or approach, though, it is important to have a direction in which to apply this model. For the purposes of this paper, we decided to focus our efforts on analyzing a word root-based hypothesis of transfer. In terms of the original LFA operations, words are themselves atomic skills. We allow merges of words on only those skills which have the same word root, according to the Porter stemming algorithm [10]. For example, consider Figure 1, which depicts merging 'cat' and 'cats'.

Student	Skill	Skill Opportunities
A	cat	1
A	dog	1
A	cats	1
A	cat	2
B	platypus	1
B	cats	1
B	cat	1

Student	Skill	Skill Opportunities
A	cat*cats	1
A	dog	1
A	cat*cats	2
A	cat*cats	3
B	platypus	1
B	cat*cats	1
B	cat*cats	2

Figure 1. Left: Original Data, Right: Data after Merging 'cat' and 'cats'

Note that 'Skill Opportunities' denotes the number of times that that particular student has seen the given word at that point in time, as the entries are sorted by student, then by timestamp. From the learning curves estimated from each of these tables, a better fit on the second model would suggest that there is evidence of transfer between 'cat' and 'cats'. We can make this observation because modeling both of them as the same word made our model of student performance, and presumably our model of the students' representations, better. It is worth noting, from an implementation standpoint, that the data occasionally had instances where the Reading Tutor read the word to the student before he had a chance to read it. Since it was in fact an opportunity for the student to see the word, we made the decision to count it towards the Skill Opportunities. However, we declined to consider it a success or failure, since such instances were not a true test of the student's performance.

Our goal is to determine whether we can achieve a reliable improvement in the cognitive model by performing some subset of the allowable merges. If so, then we would have evidence of a correlation between the word root representation of words, and the student representation. The philosophical implication of the final set of skills chosen is that this model best represents how a student represents the task at hand. If the behavior of the student is better fit by the final model than before the operators were applied, we assume this is consistent with having discovered a better approximation to their mental representation.

4. Model and Implementation

Now that we have defined a dataset and a goal, the next crucial step is to decide how to represent it in terms of the logistic model in LFA. The original logistic model [1] is as follows:

$$\ln\left(\frac{p}{1-p}\right) = \sum \alpha_i X_i + \sum \beta_j X_j + \sum \gamma_j Y_j T_{ijt}$$

Equation 1

p = the probability to get an item right
X = the covariates for students
Y = the covariates for skills
T = the covariates for the number of opportunities practiced on the skills
Y T = the covariates for interaction between skill and the number of practice opportunities for that skill
α = the coefficient for each student, i.e. the student's "smarts"
β = the coefficient for each rule, i.e. the skill's difficulty
γ = the coefficient for the interaction between a skill and its opportunities, i.e. the learning rate

Note that there is one parameter for each student, as well as two parameters for each skill. For the dataset listed above this model would result in well over 1,000 parameters, which is bordering on infeasible already, before accounting for the vast number of instances to be handled. While we could theoretically compute model fit statistics for 650,000 instances of input as in the original LFA code, time and memory restrictions quickly inhibit the performance of even the most advanced of statistical software packages. While initial attempts were made by the authors to render this problem tractable, it became apparent that developing an alternative approximate solution might yield reasonable results. The first aspect of the model we focused on was the number of student parameters. These student variables serve as indicator variables in the original regression, identifying exactly which instances belong to each student. Rather than treating student identity as a separate factor, which would result in a number of parameters equal to the number of students, we instead computed the proportion of words accepted as correct in our dataset, for each student. This value is then used as a covariate in our model, requiring only a single parameter to represent an arbitrary number of students. While this simplification loses the representational power that the original model had regarding the students, it makes for a much more computationally feasible regression.

Having reduced the number of parameters significantly from the student perspective, we then focused on developing an approximation for the other part of the model; namely, those parameters relating to the skills. The immediate problem at hand is that of tractability; LFA recalculates the entire model after each application of an operator. As mentioned earlier, we are focusing on analyzing the word root model of decoding representation, which means that each transition involves a merge of a pair of related words. However, this focus implies that (now that the specific student parameters are gone) much of the data has no impact on a merge. For example, in the merge detailed in Figure 1, only those instances with 'cat' or 'cats' would need to be considered in the regression. A considerable savings in both time and memory required follows from the observation that only a small portion of the data will involve the two words to be merged. These simplifications also allow us to consider all merges to be independent, excepting merges with the same word root. From an implementation standpoint, proper caching of each independent merge in the best-first search tree can lead to greatly improved run times.

These smaller models introduce an element of approximation, especially due to the student smarts parameter, as seen in Equation 1. However, the skill coefficients obtained through logistic regression for each model can still be interpreted similarly to the original LFA ones, as can the model fit. Additional empirical validation came from a comparison of the decisions made by our regression model on the Geometry Cognitive Tutor data used in [1], where we saw similar trends to the published results.

The last topic related to defining our approach to LFA is to actually decide on some heuristic by which to score an individual model, in order to allow comparison between models. Numerous scoring metrics have already been proposed for LFA, including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), log likelihood of the data, as well as the coefficient of determination (R-squared).

An in-depth discussion of the pros and cons of each of the criteria, as well as their definitions, is beyond the scope of this paper, for which the authors refer you to [11]. However, for the purposes of this paper we chose to direct our searches using AIC [12]. Since we are working with a large dataset, we wanted to minimize the effect of numerous instances upon the score, unlike BIC, which tries to factor in this effect. Additionally, the coefficient of determination was ruled out due to the lack of agreement by experts as to its usefulness in the logistic model [13]. A 'good' merge (i.e. one that should be allowed) is defined to be one where the AIC score of the model improved when the words are merged.

5. Experiments and Results

Our goal for this framework is to observe what sort of relationship exists between the word root model and student mental representations. The first question we must ask is: "Does the word root model actually provide us with a better model of the mental representation we are trying to examine?" In order to answer this question we examined each pair of skills in our model, to determine how many benefit from merging, and to what extent. The results can be seen in Figure 2. Points above the 0 line indicate that there was a positive benefit from performing the merge on the corresponding skill. We immediately recognize that there seem to be some skills that benefit greatly from merging, whereas others actually made the model score worse. One notable fact from this curve is that approximately 113 out of 177 of the skills would be merged (64%), if we were to choose the best-fit model for the population. Since each pair of skills is independent, we can also compute the net effect of these merges. Overall, there was an improvement of 581 points to the AIC score for the entire model, as compared to the total AIC score of 93,697, summed over the skills which benefit from merging.

In order to observe relative improvements to the students' representation due to the use of the word root model, we grouped students into subsets which we hypothesized would lead to different degrees of transfer. We felt that the word identification component of the Woodcock Reading Mastery Test [14], designed to assess a student's ability to read words of varying difficulty, would be a productive way to separate students. We divided students in two ways. First, we separated students into three equal-size groups based on their pre-test performance. Second, we created three equal-size groups based on the student gains from pre- to post-test. The first split reflects the impact of student knowledge on mental representation; the second split reflects whether the students' rates of learning and mental representations are intertwined. Perhaps students learn more quickly because they have a more efficient representation of the domain?

With these two divisions in mind, our purpose is to determine whether allowing merges of words with the same root would provide a better model fit. In order to measure this effect, we focus on two related experiments. Given the models for each pair of words and their resulting merge, we first examine the relationships between model score (AIC) and the skills being merged, over each of the divisions. These relationships provide us with two interesting pieces of information. First, within each of the three individual groups, the corresponding relationship provides some intuition as to what effect the merge has on the model representation itself. The perspective of such a result tells us whether there is any benefit to considering the word root model as a potential improvement to students' mental representations. Second, the relationships between each of the three curves provide information concerning the students themselves. For instance, we can ask ourselves "Who do we think is more likely to be using a word root representation of reading: a student who performs poorly on the pre-test, or one who did well?"

The second of our experiments involves looking directly at the learning rates for each of the skills. In LFA, the coefficients involving the learning rates and difficulty for skills are crucial for understanding how students represent words. In our pair-wise model, we produce similar parameters, and would like the ability to make statements concerning their interpretation. To this end, we look directly at the slope parameters for each pair of skills before and after merging, and attempt to discern whether there is any noticeable relationship between their initial values and the merged one. We can then examine this relationship to make claims about the transfer effect between words with the same root.

5.1. Results for Experiment 1: Model Scoring Comparisons

As previously described, we measured the model score value (AIC) for each pair of words with the same word root. For each pair, LFA resulted with two values: one score for the pair prior to merging, and one after having performed the merge. Figures 2 and 3 demonstrate the relationships between the three groups in each division, where the dependent variable on the y-axis is the improvement in model score for each pair of words merged.

For the pre-test divisions shown in Figure 2, we note that there seems to be a reliable difference between the high performing students and the low performing ones. Namely, each group has some words which, upon merging, improve the model score. However, the high performance students have the greatest proportion of these. From another perspective, there also seems to be an opposite trend with the words which do poorly when merged. The degree to which merging hurts the score is much less on the high proficiency students. In contrast, there is no such discernable pattern in the student gain graph in Figure 3. To better visualize how the groups compared, Table 1 contains the percentage of words that resulting in an improvement in AIC.

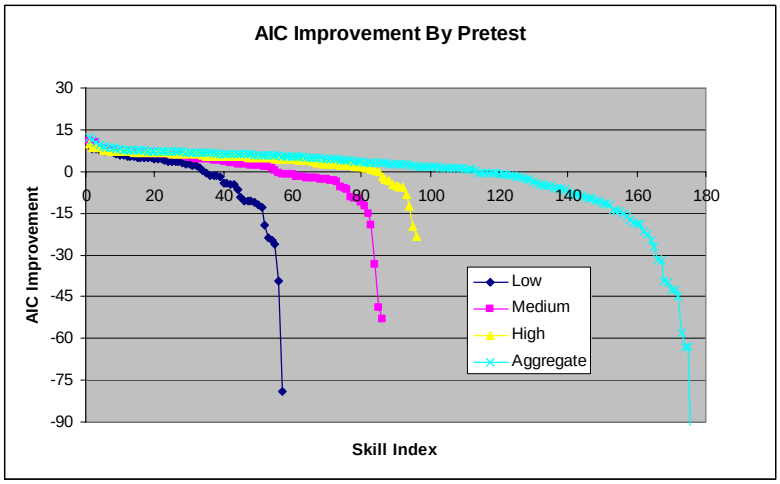

Figure 2. Impact of Merging for Pre-test (with overall aggregate)

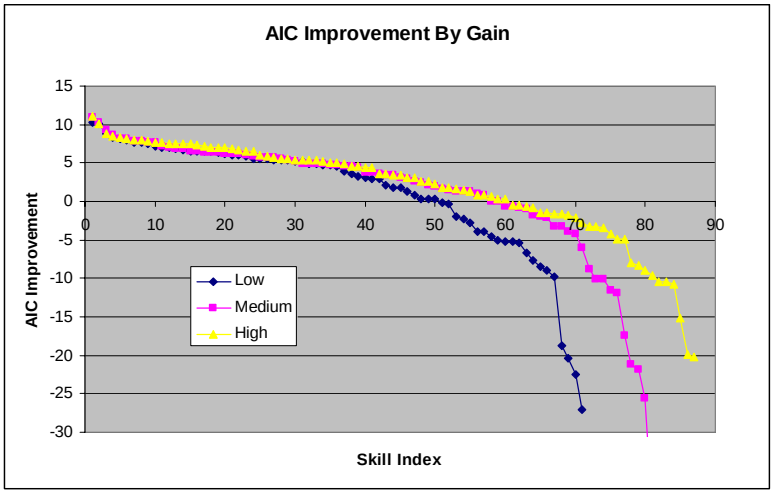

Figure 3. Impact of Merging for Gain

Upon performing a χ^2 test of reliability on the results in Table 1, we found that the only two reliable relationships present in the graph were in the pre-test group between the high proficiency students and each of the other two groups. High proficiency students showed signs of transfer for 89% of the words, while low and medium proficiency groups only showed signs of transfer for 59% and 64%, respectively. In order to consider these results at a finer grain, we performed additional reliability testing in the form of statistical bootstrapping. Bootstrapping is a resampling approach made popular by Efron in 1979 [15]. Among the approaches several desirable traits is a complete independence from assumptions concerning the distribution of the population, which is unknown for our application. In this case, our population is the set of differences in AIC score for each pair of student divisions. We bootstrapped 20,000 examples, from which we determined that there was only a reliable difference (with 95% confidence) between the high proficiency pre-test students and the other two pre-test groups. The resulting p-values are summarized in Table 2. This reaffirms the initial results from the χ^2 test.

From these observations we can conclude that there exists a reliable difference between these student divisions, using pre-test scores as an indicator. Specifically, high proficiency students exhibit more transfer at the word root level. However, using student gains over the year is not useful as a predictor.

Table 1. Percent of possible merges that improve AIC score

Student Group	Test Score Measure	
	Pretest	Gains
Low	59%	70%
Medium	64%	71%
High	89%	69%

Table 2. P-values for bootstrapping comparisons between divisions

Student Group	Test Score Measure	
	Pretest	Gains
Low vs. Medium	.18	.52
Medium vs. High	.01	.28
Low vs. High	<.002	.68

5.2) Results for Experiment 2: Learning Rate Relationships

An alternate way of exploring the models developed by LFA is to focus on the learning rate for each of the skills. Specifically, prior to merging two skills such as 'cat' and 'cats', the regression model has two different parameters, corresponding to the last summation in Equation 1, which represent the learning rates for each skill. We can similarly analyze the learning rate obtained from the merged skill. We take the average of the initial learning rates for each pair of skills and regress the learning rate of the final merged skill against it. In theory, the more positive the resulting slope is, the more transfer is occurring. More importantly, we can compare these values within both sets of divisions, to determine what types of students seem to get the most transfer from the word root model. In order to maintain our goal of computing interpretable results, we decided to remove any data points which had values greater than five times the value of any other data points for a particular regression. This definition of outlier resulted in the removal of one data point from the high proficiency pre-test division, and two data points from the high proficiency gains division. Our results are summarized in Table 3. In general, this ratio represents the fact that if a skill has very little associated transfer, we can expect that the merged skill will have a lower learning rate than if the word had exhibited a high degree of transfer.

Table 3. Degree of transfer from words to word roots

Student Group	Test Score Measure	
	Pretest	Gains
Low	.30	.28
Medium	.26	.43
High	.40	.41

Table 4. P-values for bootstrapping comparisions between regression slopes

Student Group	Test Score Measure	
	Pretest	Gains
Low vs. Medium	.78	.11
Medium vs. High	.17	.97
Low vs. High	.55	.27

For the pre-test divisions, we first note that the degree of transfer doesn't necessarily increase steadily as our estimate of initial student knowledge increases. However, the high proficiency pre-test students seem to have the strongest positive relationship, which implies that we observed the greatest degree of transfer from their data.

When examining the gains divisions, we have the opposite problem with the middle group, as they seem to be exhibiting approximately as much transfer as the high gain students. Nevertheless, it appears as if the low gains

students are exhibiting the least degree of transfer, which supports our overall conjecture that stronger students will have a stronger degree of word root transfer than weaker students. We note that these results paint a slightly different picture than the previous experiment, which could make more substantial claims about the pre-test results while providing little information about the student gains. However, one consistent result is that high proficiency students exhibit the greatest degree of transfer.

Similarly to our previous experiment, we performed statistical bootstrapping to determine to what degree these results are reliable. Our resulting p-values, summarized in Table 4, indicate that we cannot reliably depend on the noted trends, perhaps due to a great deal of variation in the slope data.

It is worth noting that these overall results are generally harmonious with those in [16] on the same dataset, especially with respect to higher proficiency students demonstrating a greater degree of transfer between words with the same root. It remains a matter of contention concerning the students in the medium proficiency group, since no results seem to indicate anything with certainty. However, all results seem to indicate a difference between the degrees of transfer in low and high proficiency students.

6. Contributions, Future Work, and Conclusions

In this paper, we have introduced the LFA framework to a new domain: student modelling in reading. Our pair-wise approximation to the original model for LFA allows for tractable computation on datasets orders of magnitude greater than those used with the original code base[1]. We have used these new capabilities of LFA to demonstrate that the use of word root representation improves our model of student knowledge. Furthermore, the experiments in this paper imply that there is a difference in students' mental representations based on two alternate metrics: first, their initial knowledge at the beginning of the school year, and second, their improvement over the course of the year. We have found a positive correlation between pre-test scores and the proportion of word merges that are predicted to improve the fit of our model. Additionally, we have explored the relationships between learning rates in each division of students. From these relationships we demonstrated that students who perform well on the pre-test seem to exhibit a greater degree of transfer than the other groups. On the other hand, students who are deemed to have gained little over the course of the year (unreliably) exhibit a rather low degree of transfer. Furthermore, we have introduced the application of two metrics for evaluating the models in LFA; namely, by examining the percentage of merges that occur, as well as examining the relationship between learning rates to approximate the amount of transfer.

While our model has provided useful preliminary results, many directions of exploration still remain. To begin with, an analysis or classification of which words were beneficial to merge, both within and across groups, could lead to interesting observations about transfer. Also, word root is only one of many possible linguistic models of word comparison. For instance, past studies have been done with student models involving grapheme to phoneme mappings [17]. Other potential models include examining the onset or rime of each word. Even these ideas only scratch the surface of possible models of reading representation; it is much more intuitive that each student's representation is comprised of an amalgamation of these models. This will also lead to more complex applications where the factors are potentially dependent on each other, making the composition of operations a nontrivial problem.

From an implementation standpoint, there exist a number of optimizations that could be considered for future iterations. Among these include an improvement to the underlying model search, where each state is represented not by an additional copy of the data, but by a set of operators performed to reach the state. This optimization would drastically improve memory management for LFA when used on large datasets, and would require a relatively small amount of computation to apply all of the operators to generate the data when necessary, even if the A* search were very deep. Also, preliminary research was done into locating a software package that could handle a larger dataset using the full LFA model [18]. While performing the entire search using this model might still be intractable, it seems reasonable to believe that a hybrid approach might be taken which combines the faster approximation presented here, and the extensive evaluation offered by full LFA.

This research represents a significant step forward in developing the Learning Factors Analysis framework. By demonstrating that the approach can be approximated in the domain of reading, we have developed new intuitions concerning student mental representations as well as the effects of student proficiency. Indeed, if dividing empirical data yields different models for each group of students, perhaps maintaining a single canonical domain model is not necessarily an appropriate practice for intelligent tutoring systems. These results impact the educational data mining community by showing that LFA is a tool that can be feasibly modified to model extensive and complex corpora of data. In addition, it is capable of capturing the effects of transfer in reading, as demonstrated by the two experiments in this paper.

[1] Source code for our implementation of LFA is available online at http://www.educationaldatamining.org under "Resources."

Acknowledgements

This work was supported by the National Science Foundation, ITR/IERI Grant No. REC-0326153. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additionally, we thank Hao Cen, Brian Junker, and Becky Kennedy for their help and advice in the development of this approach.

References

1. Cen, H., K. Koedinger, and B. Junker, *Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement.* Proceedings of Intelligent Tutoring Systems, 2006.
2. Newell, A. and H.A. Simon, *Human Problem Solving*. 1972, Englewood Cliffs, NJ: Prentice-Hall.
3. Heffernan, N.T. and K. Koedinger, *A developmental model for algebra symbolization: The results of a difficulty factors assessment.*, in *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. 1998, Lawrence Erlbaum Associates, Inc: Mahweh, NJ. p. 484-489.
4. Croteau, E.A., N.T. Heffernan, and K. Koedinger. *Why Are Algebra Word Problems Difficult? Using Tutorial Log Files and the Power Law of Learning to Select the Best Fitting Cognitive Model*. in *7th International Conference on Intelligent Tutoring Systems*. 2004. p. 240-250: Springer Berlin.
5. Mostow, J. and G. Aist, *Evaluating tutors that listen: An overview of Project LISTEN*, in *Smart Machines in Education*, P. Feltovich, Editor. 2001, MIT/AAAI Press: Menlo Park, CA. p. 169-234.
6. Mostow, J., J. Beck, R. Chalasani, A. Cuneo, and P. Jia. *Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach*. in *Fourth IEEE International Conference on Multimodal Interfaces*. 2002. p. Pittsburgh, PA.
7. Banerjee, S., J. Mostow, J. Beck, and W. Tam. *Improving Language Models by Learning from Speech Recognition Errors in a Reading Tutor that Listens*. in *Second International Conference on Applied Artificial Intelligence*. 2003. p.
8. Beck, J. *Using learning decomposition to analyze student fluency development*. in *ITS 2006 Educational Data Mining Workshop*. 2006. p. Jhongli, Taiwan.
9. Freyberger, J., N.T. Heffernan, and C. Ruiz, *Using Association Rules to Guide a Search for Best Fitting Transfer Models of Student Learning*. 2004, Worcester Polytechnic Institute: Worcester, MA. p. 10.
10. van Rijsbergen, C.J., S.E. Robertson, and M.F. Porter, *New models in probabilistic information retrieval*. 1980, British Library: London.
11. Cen, H., K. Koedinger, and B. Junker, *Automating Cognitive Model Improvement by A* Search and Logistic Regression.* Proceedings of AAAI 2005 Educational Data Mining Workshop, 2005.
12. Akaike, H., *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 1974. **19**(6): p. 716-723.
13. Meynard, J., *Coefficients of determination for multiple logistic regression analysis.* American Statistician, 2000. **54**: p. 17-24.
14. Woodcock, R.W., *Woodcock Reading Mastery Tests- Revised (WRMT-R/NU)*. 1998, Circle Pines, Minnesota: American Guidance Service.
15. Efron, B., *Bootstrap Methods: Another Look At The Jackknife.* The Annals of Statistics, 1979. **7**(1): p. 1-26.
16. Zhang, X., J. Mostow, and J. Beck, *All in the (word) family: Estimating transfer from reading similar words.* Educational Data Mining Workshop, (under review).
17. Chang, K., J. Beck, J. Mostow, and A.T. Corbett. *Using speech recognition to evaluate two student models for a reading tutor*. in *12th International Conference on Artificial Intelligence in Education*. 2005. p. 12-21 Amsterdam.
18. *SPLUS 7.0 Enterprise Developer*. 2005, Insightful Corp.: Seattle, WA.