Learning Factors Analysis Learns to Read

James M. Leszczenski

CMU-CS-07-136
August 2007

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Joseph E. Beck, Chair
Vincent Aleven

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science.*

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.

Keywords: Learning Factors Analysis, Reading Transfer, Intelligent Tutoring Systems, Project LISTEN

Abstract

Learning Factors Analysis (LFA) has been proposed as a generic solution to evaluate and compare cognitive models of learning [Cen et al., 2006]. By performing a heuristic search over a space of cognitive models, the researcher may evaluate different representations of a set of skills. This search, however, is computationally intractable for large datasets. 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 that permits transfer at the level of word roots better reflects actual student learning data. 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. 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. Additionally, we provide insight into developing metrics designed to classify in advance whether particular operations within LFA will exhibit transfer.

Acknowledgments

Thanks to both Brian Junker and Becky Kennedy for their help and advice in the development of this approach. Thanks also to Hao Cen, who provided helpful comments on earlier versions of this work. Special thanks to Vincent Aleven, for providing me with a different perspective and excellent comments as a member of my committee. Finally to my advisor, Joseph Beck, for ideas, insight, and pointed humor that were invaluable to my progress.

Contents

1 **Introduction** **1**

1.1 Defining a Problem 1

1.2 Identifying Transfer within a Domain 3

1.3 Domain Instantiation 5

2 **Learning Factors Analysis: Infrastructure and Approach** **9**

2.1 The Original Statistical Model 10

2.2 Choosing a Transition Function 13

2.3 Designing a Faster, Approximate Model 14

2.4 Searching the Model Space 18

3 **Experiments and Results** **21**

3.1 Identifying Main/Aggregate Effects through Transfer and Model Improvement 21

3.2 Analysis through Student Proficiency Disaggregation 24

3.3 Predicting What Skills Demonstrate Transfer 30

4 **Contributions, Future Work, and Conclusions** **37**

4.1 Contributions 37

4.2 Future Work 38

4.3 Conclusions 40

Bibliography **43**

List of Figures

2.1 The Original Logistic Equation for LFA 11
2.2 Left: Original Data, Right: Data after Merging ‘cat’ and ‘cats’ 14
2.3 The New Approximate Logistic Model Equation, Before Merging 16

3.1 Aggregate Impact of Merging for AIC 22
3.2 Aggregate Impact of Merging for BIC 23
3.3 Disaggregate Impact of Merging by Pre-test Divisions for AIC 25
3.4 Disaggregate Impact of Merging by Pre-test Divisions for BIC 26

List of Tables

3.1 Percent of Possible Merges that Demonstrate Transfer 27

3.2 Word Feature Classification on Aggregate Data, AIC 33

3.3 Word Feature Classification on Aggregate Data, BIC 33

3.4 Model Feature Classification on Aggregate Data, AIC 34

3.5 Model Feature Classification on Aggregate Data, BIC 34

Chapter 1

Introduction

1.1 Defining a Problem

The task of modeling 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 [Newell and Simon, 1972]. 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 [Heffernan and Koedinger, 1998]. 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 [Croteau et al., 2004].

We cannot, however, simply decide that 'modeling student cognition' is our goal. There are countless ways of approximating this cognition, and similarly many ways of evaluating how effective the approximations are. Instead, we decided to instantiate the problem with a more specific focus, by looking at our cognitive models in terms of cognitive transfer.

Before explaining how to analyze transfer, or even recognize it, we must first provide a baseline definition of transfer, after which we can design a methodology by which to systematically recognize it, and apply it to our modeling. In one of the premier texts on cognitive transfer [Singley and Anderson, 1989], the following definition is offered:

"The study of transfer is the study of how knowledge acquired in one situation applies (or fails to apply) in other situations."

While Singley and Anderson refer to transfer in terms of 'tasks,' this definition is particularly relevant in terms of the basic knowledge components in a cognitive model. Consider any two skills (or, equivalently, tasks) that a model attempts to represent. Unless we believe that these skills have nothing in common, then it is possible that some subset of their underlying knowledge components overlap. Thus, transfer from one of these skills to the other can potentially be observed, if practice on the knowledge components involved with one are sufficient to improve performance on the other skill, due to these commonalities.

1.2 Identifying Transfer within a Domain

While we have offered some basic insight into transfer in cognitive modeling, the practical application of this observation requires that such cognitive transfer be palpable. Fortunately, the advent of computing and its uses in the advancement of education though cognitive modeling brought about the introduction of intelligent tutoring systems (ITS), which allow for a more data-driven approach to be taken. Within such a system, there will generally be a single expert-developed model of how the student is expected to learn the task at hand. Such a model could be as simple as prompting until the student makes a proper response, or as complex as involving model tracing and knowledge tracing to try and maintain an accurate concept of the student's mental state [Koedinger and Anderson, 1997]. Either way, the data collected from such a tutor is a rich source of knowledge about actual student cognitive models. This student performance data can then be used as an attempt to identify the knowledge components underlying the skills tested in the cognitive model.

We shall address the issue of recognizing transfer in a model in terms of our particular instantiation of the transfer problem. Learning Factors Analysis (LFA) has been proposed as a generic solution to evaluate and compare many potential cognitive models of learning [Cen et al., 2006]. As input, LFA takes an initial statistical 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 [Cen et al., 2006], LFA can locate skills that contain such commonalities.

From the perspective of a student's representation, the research task is to approximate how and where transfer occurs. The LFA framework attempts to model transfer between skills under the assumption that if two skills A and B are better modeled as a single combined skill rather than individual skills, then this will equivalently indicate that practice on either A or B will transfer to the other. This in turn 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 learning 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 with similar probability 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.

We will return to and expand upon each of these ideas in turn. However, now that we have developed the basics behind identifying transfer, we must present our domain-specific information more formally.

1.3 Domain Instantiation

In order to guide our efforts in extending and improving LFA as a tool for exploring student cognitive model improvement, we required a test bed containing an appropriate corpus of data. We selected our data from the Project LISTEN Reading Tutor in the 2003-2004 school year. The Reading Tutor [Mostow and Aist, 2001] is a type of intelligent tutoring system (ITS), which assists children 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 [Mostow et al., 2002]. 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, generally between the ages of 5 and 12, 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% [Banerjee et al., 2003]. 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 [Beck, 2006], 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 some of the biasing 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.

Chapter 2

Learning Factors Analysis: Infrastructure and Approach

Learning Factors Analysis (LFA) has been proposed as a method by which to “combine statistics, human expertise and combinatorial search to evaluate and improve a cognitive model” [Cen et al., 2006]. The LFA framework has three functional aspects.

First, since our goal is to attempt representation of student cognition, it must rely on some data-driven approximation to the students and their performance. We describe the use of a logistic regression model as a statistical approximation in the following subsection “The Original Statistical Model,” which also includes a rationale for this choice of model.

The second functional aspect of LFA is the expert-defined transition function from one state/model to another. Referred to as ‘difficulty factors’ in [Cen et al., 2006], these are intuitively a way of defining the entire legal model space for the domain. In “Choosing a

Transition Function" we discuss at a higher level the basic idea behind what these factors can accomplish, and how we specifically adapted them to our domain-specific task. We proceed to introduce a new transition-based statistical model in the subsequent section, designed as a fast approximation to the original. Note that the new model we propose is domain-independent, and has no dependency on the specific transitions we describe.

The final aspect of LFA is the combinatorial search through the space of all models defined by the difficulty factors described. This search will be described in the subsection "Searching the Model Space".

2.1 The Original Statistical Model

The original logistic model is listed in Equation 2.1. In this model, we define the parameters as follows. 'p' represents the probability that an item is correct. The capital letters represent the variables in our logistic model. X represents the set of variables denoting the students. Y represents the set of variables denoting the skills performed. T represents the variables for the number of skill practice opportunities, of which we have one variable for each skill. Hence, 'YT' denotes the interaction between a skill and its number of practice opportunities. The Greek letters denote the coefficients learned by the logistic model on each variable. The α coefficients, for each student, are designed to represent individual student proficiency, or 'smarts.' The β coefficients are designed to represent individual skill difficulty relative to the other skills. Finally, the γ coefficients represent the learning rates for each of the skills, learned by the logistic model.

The assumptions behind this original model can be found in [Cen et al., 2006]. This

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

Figure 2.1: The Original Logistic Equation for LFA

model is conceptually similar to item response theory, in the sense that it uses a logistic shape to map student characteristics to a type of probabilistic performance response. However, LFA goes a step further and does not make the assumption that student knowledge is static. Instead, it allows modeling student improvement through practice. The model itself is data-driven, and based upon a set of 0/1 training data including time-ordered triples of students, skills practiced, and the logged success or failure for each of these instances. There are three separate types of variables. We have one variable for each student, which is designed to account for the fact that each student will be operating at his own level of proficiency. The second set of variables relates directly to the skills being practiced, or in the Reading Tutor instantiation, the words being read. We note that for our domain, the terms ‘word’ and ‘skill’ are nearly synonymous. The only difference lies in the fact that we use the conventional definition for ‘word,’ whereas a skill can represent a more generic construct. This construct may consist of a single word, or it may consist of a set of words, as we shall later describe. There is one variable for each of these skills, which is designed to represent the difficulty of that particular skill. Finally, the third set of variables is designed to represent the learning rate associated with each of the skills.

For a comprehensive analysis of the model, we also include complexity, since tractability is a crucial practical concern. As described, for a regression model with X students and Y skills, the number of parameters is $X + 2Y$. This alone may initially seem reasonable, in order to get a proper view of a cognitive model for the entire population in which we are interested. However, this is merely the number of parameters of the model; it doesn't take into account the asymptotic running time of a logistic regression model, which depends both upon the number of parameters as well as the number of instances upon which the model is trained.

Logistic regression as a classification technique is a well-researched method, and has become one of the core tools used in data mining and related fields [Komarek and Moore, 2005]. As such, a number of implementations, each with their own set of assumptions, optimizations, and features have been developed. However, since our data have no guarantees regarding sparsity or other features, it is reasonable to expect no better running time than a basic implementation [Komarek and Moore, 2003]. Our implementation relied on an iteratively re-weighted least squares (IRLS) method within the framework of R, an open source statistical package [Ihaka and Gentleman, 1996]. IRLS logistic regression is generally run for a constant number of iterations or until convergence on a maximum likelihood estimate of the best model. Even in this second case, the number of iterations is empirically found to be rarely more than a small constant. Each update in a basic straightforward implementation, however, requires O($M^3 + MR$) time, where M is the number of parameters in the model, and R is the number of instances of data [Komarek, 2004]. If we are willing to consider an approximation using conjugate gradients then , then the computation of a single iteration can be reduced to the number of non-zero entries in the matrix representing the value of every parameter for every instance [Shewchuck, 1994]. An upper bound on this value is O(MR). While much improved from the base implementation, it is still clear that logistic regression is a computationally intensive procedure on all but the smallest datasets.

Since this scaling issue would be rather impractical for the Reading Tutor dataset as described earlier, we considered ways in which to minimize the size of the model, while attempting to maintain as close an approximation to the effects of LFA as possible. We shall return to this intuition after introducing model transitions, which make simpler the conceptualization of our novel approximate model.

2.2 Choosing a Transition Function

Specifically, prior work on LFA [Cen et al., 2006] has proposed that models can be transformed via 'split', 'add', and 'merge' operators which generally act on a single skill at a time, making incremental transitions between similar models. In general, expert-defined intrinsic factors pertaining to observed skills are encoded along with the input. In order to allow a proper combinatorial search, starting from the model created from the initial data, these factors define changes that can be made to the current model in order to derive a new one. Essentially, the transitional operators are relabeling the instances of data to form a new dataset, derived from the original. 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 [van Rijsbergen et al., 1980]. For example, consider Figure 2.2, which depicts merging 'cat' and 'cats'. This illustrates two students who read a series of words, listed in the 'Skill' column. The left side denotes the number of practice opportunities on each word

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

Student	Skill	Practice 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 2.2: Left: Original Data, Right: Data after Merging ‘cat’ and ‘cats’

if we assume that every word is atomic, and has no effect on any other. The right side denotes the number of practice opportunities if we consider the hypothesis that the words ‘cat’ and ‘cats’ are similar enough that practicing one of them is equivalent to practicing the other, resulting in a single skill we label ‘cat*cats’.

2.3 Designing a Faster, Approximate Model

As described in the preceding two sections, the original approach to LFA required that the entire statistical model be recomputed after every operation. This model, as described Equation 2.1, requires a complex logistic regression computation, as we earlier analyzed. While this is sufficient for reasonable small datasets with low dimensionality, it is intractable for the Reading Tutor dataset, or anything with a similar order of magnitude.

Since reducing the complexity of logistic regression or reducing the size of the dataset are not reasonable alternatives, we are left with one remaining dimension to reduce: the number of parameters in the logistic regression model.

We noted that there are three types of variables in the LFA model. We shall focus first on the student variables. The original purpose of these variables in the model was to act as identifiers which mapped students to their instances in the dataset. This allowed for the computation of initial knowledge levels for each student, or individual student smarts as we called it in Equation 2.1. If we wished to focus on student performance as a primary factor in our results, this level of granularity would be useful. However, as we wish to mainly explore transfer between words in reading, the student variables in the model seem like a reasonable starting place for our approximation.

Regarding student variability, the assumption has already been made that students learn at the same rate. Hence, we want to model initial student knowledge level, or "smarts." We propose the use of an aggregate student smarts variable, rather than individual ones for each student. As an approximation, we computed the proportion of words accepted as correct in our dataset, for each student. This value was used for our reduced dataset, so that each instance of data was related to a "smarts" proportion, rather than a particular student. The rationale behind using this value is that we are still allowing our model to take into account student knowledge variance, while not modeling specific, individual variance. Thus, if we had X students in our dataset, we still only need a single variable in the logistic regression model to approximate the difference between student smarts, rather than X parameters as did the original LFA model. We assume that these approximate smarts proportions remain constant throughout the course of our modeling. Hence, we

$$\ln\left(\frac{p}{1-p}\right) = \alpha X + \beta Y_{1,2} + \gamma_1 Y_1 T_1 + \gamma_2 Y_2 T_2$$

Figure 2.3: The New Approximate Logistic Model Equation, Before Merging

avoid the recalculation of parameters for every student.

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 still that of tractability; LFA was originally designed to recalculate the entire model after each transition, which requires using the entire dataset to perform the regression. And while we have significantly reduced the model with respect to the student variables, we still have another set of parameters whose size is twice the number of skills. 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. This focus implies that (now that the specific student parameters are gone) much of the data, and hence many of the skill variables, remains unaffected by a merge.

To this end, we developed a transition-based model. Each transition is characterized by a 'before' and an 'after' state. That is, the model prior to the merge being performed, and the model after the merge has occurred. In the original model, these would each be a full computation of the model, representing our cognitive world. Now, our reduced model would merely attempt to approximate the local effect of the transition, by measuring the change in model score (which we shall introduce in the next section) between the 'before' and 'after' states.

The ‘before’ state needs only one variable for the students, one variable to represent the relative skill difficulty, and two variables to represent the learning rates for each skill prior to the merge, for a total of 4 parameters. We illustrate this model in Equation 2.3, in a format similar to how we described the original LFA model in Equation 2.1. In this new model, we once again define the parameters as follows. ’p’ represents the probability that an item is correct. The capital letters represent the variables in our logistic model. X represents the single covariate representing aggregate student smarts, where the values it takes are the proportion of words each student were determined to get correct by the data. Y represents the variable denoting the relative difficulty difference between the two skills. T represents the variables for the number of skill practice opportunities, of which we have one variable for each of the two initial skills. Hence, ’YT’ denotes the interaction between each skill and its number of practice opportunities. The Greek letters similarly denote the coefficients learned by the logistic model on each variable. The α coefficient is designed to represent aggregate ’student smarts.’ The β coefficient is designed to represent relative skill difficulty between the skills. Finally, the γ coefficients represent the learning rates for each of the two skills, learned by the logistic model.

The ‘after’ model, on the other hand, needs only the single student variable and a variable for the learning rate for the resulting skill after merging. A difficulty parameter would not be interpretable since it would be relative to itself in such a model. Hence, the ‘after’ model only requires 2 parameters.

Since each transition is characterized by a small constant number of parameters, we can ignore the ‘M’ parameter listed in the running times listed for the original logistic model. This yields a running time of $O(R)$, where R is the number of instances in the dataset. In the case of the Reading Tutor dataset, this is an improvement of several orders of magnitude over the original model.

Note that this model is still intrinsically domain-independent; we are essentially just taking a local view of the world and focusing on a small specific set of skills, rather than globally attempting to estimate the entire model. From an asymptotic and theoretic standpoint this is an immense improvement, with respect to time and memory, compared to the original model. It remains to be seen whether our weaker model representation can be of any practical use. Our results address this question directly, as well as other more specific questions.

2.4 Searching the Model Space

The final aspect of LFA is the combinatorial search, which takes the initial statistical model and our defined set of transitions as input, and proceeds to attempt finding the best possible state in model space, according to some scoring metric. Given unlimited computing time, some form of k-fold cross validation would be the best metric of comparison, due to its resistance to bias and overfitting. However, due to the inherent complexity of this approach, it would be an infeasible metric for this application, where analysis of many models would be necessary.

Numerous scoring metrics have already been proposed for LFA, including Akaike's An 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 is beyond the scope of this paper, for which the authors refer you to [Cen et al., 2005]. However, for the purposes of this paper we chose to consider directing our searches using AIC and BIC [Akaike, 1974]. Log likelihood of the data doesn't account for the parsimony of the learned model, and thus can be prone to overfitting to the data. The coefficient of determination was ruled out due to the lack of agreement by experts as to its usefulness in the logistic model [Meynard, 2000]. Both AIC and BIC take log likelihood into account, and offer a penalty term for the complexity of the model. They differ in that the penalty of AIC is dependent only on the complexity of the model, whereas BIC more strongly penalizes a complex model as more data is available. For our domain, there is no obvious choice between these two. As such, we present results from each for the experiments where it is warranted, offering a contrast between the conclusions made by these metrics. A 'good' merge (i.e. one that should be allowed) is defined to be one where the scoring metric of the model improved when the words are merged.

Now that we have sufficiently defined a model scoring comparator, we must describe the search to be done. As in [Cen et al., 2006], we perform a best-first search over the search tree of possible models. In essence, we maintain a set of models that we have evaluated at any given point in time, and choose the next model to consider based on the scores of those we have evaluated. Intuitively, this methodology assumes that a path in the search tree that has been beneficial is more likely to become increasingly beneficial. One of the greatest potential problems with this method is the necessity to store all previously computed nodes, in order to determine the best one evaluated so far. As the amount of data increases, this becomes memory-intensive. Additional complications stem from the complexity of quickly recognizing equivalent models prior to computing them redundantly. However, it is worth noting that the concerns we address with respect to the search are all practical issues that do not have any effect upon the results, and are problems that can be solved or at least minimized with an efficient implementation.

Chapter 3

Experiments and Results

So far, we have described our changes to the LFA framework. It remains to be seen, however, whether we can observe anything substantial from our modifications. Within our domain of reading, the first significant step would simply be to acquire results of any sort from the Reading Tutor data.

3.1 Identifying Main/Aggregate Effects through Transfer and Model Improvement

Our first task is to attempt to determine whether all words are indeed independent as hypothesized by the "bag of words" theory mentioned earlier, or alternately, whether the word root model we are considering can in fact lead us to a better model of cognitive representation. The results for the main effect of our experiment are seen in Figure 3.1 for

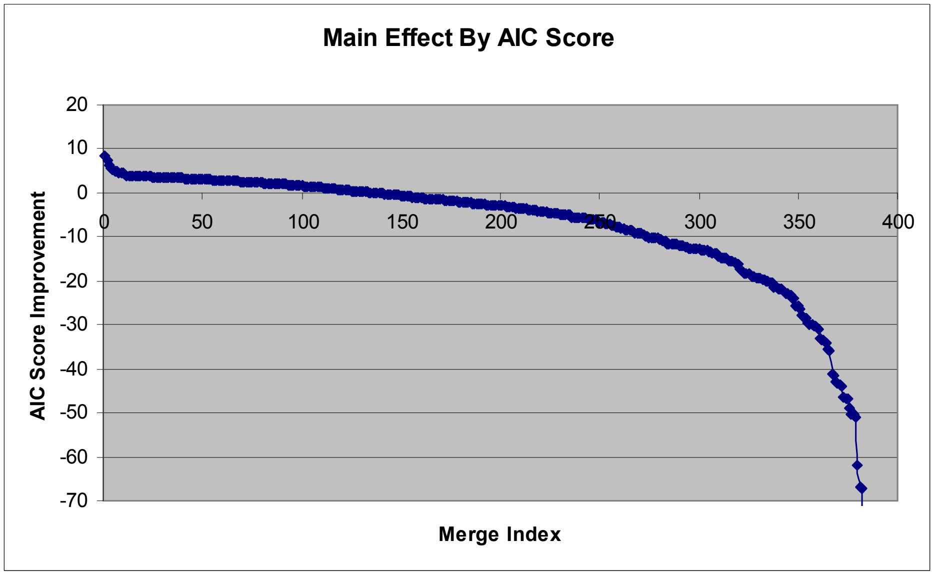

Figure 3.1: Aggregate Impact of Merging for AIC

the AIC score improvement, and Figure 3.1 for the BIC score improvement. Each index on the horizontal axis represents a pair of skills which we could potentially merge, in an attempt to improve our model. Each point above the 0 line on the vertical axis indicates that there was a benefit from performing that particular merge.

In the context of related literature, this is in some respect an attempt to examine one aspect of the dilemma of specificity of transfer. For instance, the doctrine of formal discipline states that the content being studied is irrelevant, since all topics of study exercise the mind, which is a "collection of general facilities" [Singley and Anderson, 1989]. If the doctrine of formal discipline is to be believed, then learning any word would help with the proficiency of any other word, and thus every potential merge should produce some improvement to the model. In the other direction, if we ascribe to the opposite end of the

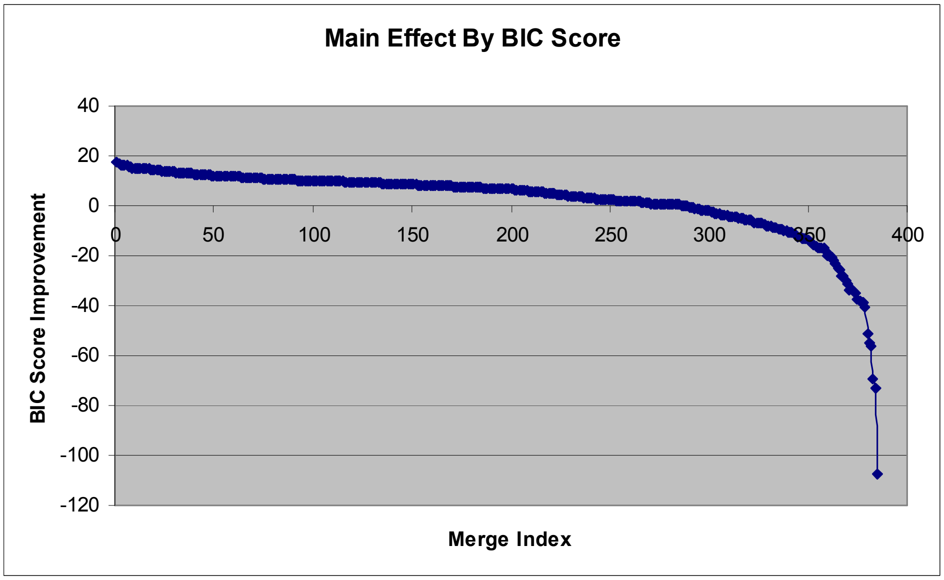

Figure 3.2: Aggregate Impact of Merging for BIC

spectrum (for example, an extremist's take on Thorndike's theory of identical elements), then we might be hard pressed to find a set of skills that would demonstrate significant transfer. From a total of 385 potential merges, 137 (35%) produced some model improvement by merging using the AIC scoring metric, whereas 287 (75%) produced some model improvement by merging them together using the BIC score. Within the framework of our experiments, it seems that something in between the two extremist theories of transfer is closer to reality.

One of the first things we note is the difference in prediction between these two metrics of model fit. In the case of BIC, an overwhelming majority of skills seem to benefit from merging, whereas in the other case, only a third of them seem to glean any benefit from the more parsimonious model. With respect to the domain in question, we have no a priori reason to believe that either scoring technique is more correct than the other. As such, our discussion and results from this point on will include both scoring metrics, and the choice of the more believable one shall be left as a future research direction.

Discussion about scoring metrics aside, these results indicate that if our statistical approximation to the student's cognitive model is to be believed, then the word root model of representation does in fact have some degree of influence on transfer in reading. This result alone lends credence to our hypothesis that we can observe the transfer effect by using a word root model. However, we can further explore this result along a new dimension, by disaggregating students into several groups.

3.2 Analysis through Student Proficiency Disaggregation

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 exhibit varying degrees of transfer. We felt that the word identification component of the Woodcock Reading Mastery Test (WRMT) [Woodcock, 1998], designed to assess a student's ability to read words of varying difficulty, would be a productive way to separate students. The WRMT is a paper-based exam. It was administered to the students by a teacher once prior to usage of the Reading Tutor as a pre-test, and then once again after the conclusion of their usage of the Tutor at the end of the school year, as a post-test. For this experiment, we divided students' data 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

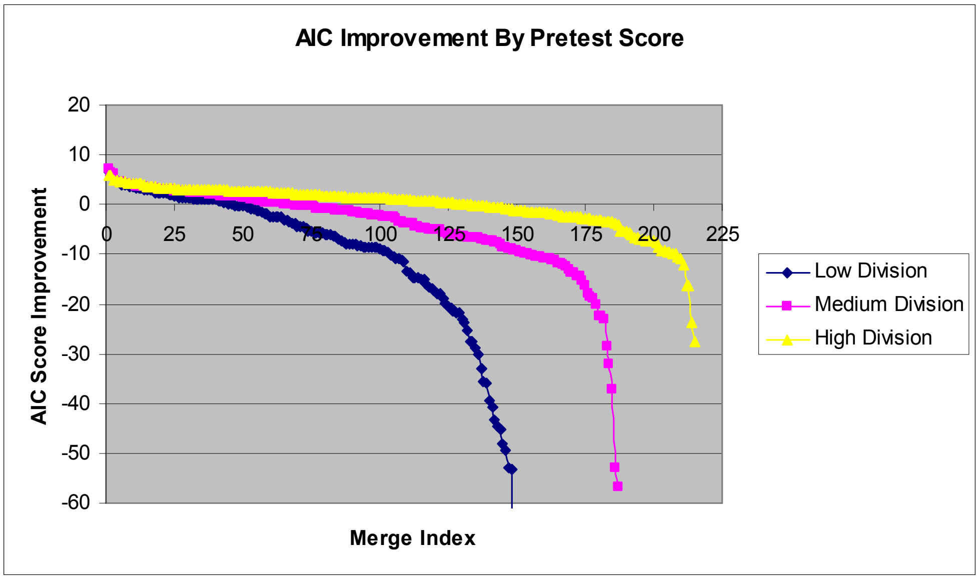

Figure 3.3: Disaggregate Impact of Merging by Pre-test Divisions for AIC

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?

Our results from the second of these divisions yielded no significant or interesting results for this and subsequent experiments, and have thus been left out of our documentation from this point forward. Figure 3.2 and Figure 3.2 represent our results for the pretest student division experiment with AIC and BIC respectively. From these visualizations, it seems clear that as the proportion of skills merged seems quite positively correlated with the pretest score division. Hence, our lack of statistically significant results from the test gains divisions is puzzling, as we would intuitively expect proficiency to be highly correlated to performance during the school year.

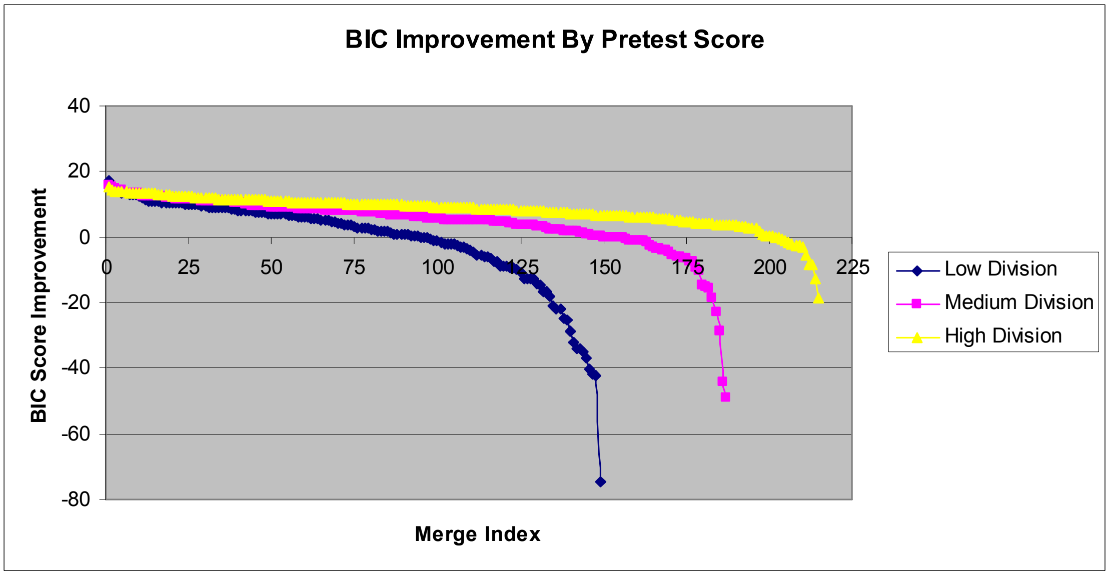

Figure 3.4: Disaggregate Impact of Merging by Pre-test Divisions for BIC

To further characterize these results, consider Table 3.1. Here, we have discretized the results from Figure 3.2 and Figure 3.2, in terms of the proportions of words merged for each scoring metric. We use the proportion rather than the absolute value since each group contains a different number of merges, after having been screened by the criteria mentioned earlier. There is a decisive upward trend for the BIC results, and the high proficiency students in the AIC grouping also seem to have far more numerous beneficial merges than do the low or medium proficiency groups.

Upon performing a χ^2 test of reliability on the results, we found that the only two reliably different relationships present in the AIC test score measure were between the high proficiency students and each of the other two groups. Using the BIC test score measure, all three pairs of student divisions was found to be reliably different.

Table 3.1: Percent of Possible Merges that Demonstrate Transfer

	Test Score Measure	
Student Group	AIC	BIC
Low	31%	64%
Medium	36%	80%
High	62%	94%

To analyze these results at yet a finer grain we proceeded to perform statistical reliability testing for these proportions in the form of statistical bootstrapping. Bootstrapping is a resampling approach made popular by Efron in 1979 [Efron, 1979]. Among the approach's 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 a reliable difference (with 95% confidence) between each of the three pairs for both AIC and BIC. More specifically, between the low and medium proficiency groups there was a statistically reliable difference with p-values of .02 and .004 for AIC and BIC respectively, while all other p-values were less than .001. Note that all p-values subsequently mentioned were generated in the same fashion.

These results indicate that initial pre-test scores are positively correlated with more compact word root models. We can conclude that there exists a reliable difference between these student divisions, using pre-test scores as an indicator. And, while each pair of divisions was determined to have a reliable difference, the numbers themselves yield the observation that high proficiency students exhibit more transfer at the word root level than medium or low proficiency students, by a large margin.

We performed one additional experiment as a potential metric for measuring the degree of transfer. 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, as seen in Equation 2, which represent the learning rates for each skill. We can similarly analyze the learning rate obtained from the merged skill. For instance, the skill ‘cat’ might have a learning rate coefficient of .46, and the skill ‘cats’ might have a learning rate coefficient of .23. Upon merging these skills, the learning rate coefficient for ‘cat*cats’ might be .32 (since the units of a logistic regression are in logits, such direct examination of coefficient values is not immediately intuitive).

We set up a linear regression model with the average of the initial learning rates for each pair of skills as our dependent variable (from our example, .345), and the learning rate of the final merged skill against it as the independent variable (from the example, .32). In theory, the more positive the resulting slope is, the more transfer is reflected by the data. 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.

Perhaps more importantly, we can compare these values within our student 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 absolute 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.

Our results for this division-based experiment are as follows. For the low proficiency group, we regressed a slope of .28. The medium proficiency group had a slope of .27, and for the high proficiency group, the slope was computed to be .38. It is especially worth noting that while the low and medium proficiency student seem to have essentially the same degree of transfer under this metric, the high proficiency student have a much higher rate. Once again, as an analysis of the reliability of these results, we bootstrapped 20,000 samples at 95% confidence. As was perhaps expected based on a quick examination of the numbers, we found that there were statistically reliable differences that existed between the high proficiency students and the other two divisions ($p<.01$), whereas there was no such difference between the lower and medium proficiency students ($p=.85$). This lends further credence to the idea that the word representation used by high proficiency students involves, to some degree, the use of word roots.

We then proceeded to take the previous experiment one step further, by recognizing that the set of skills specifically merged could be (and was) different for each of the three divisions. As such, we hypothesized that a potentially more fair approach to determining where transfer would occur would be to only consider the set of words that were merged under all three divisions. Using BIC as our metric of whether a skill was merged or not, there were 65 pairs of skills which were merged in all three divisions. One outlier was removed from the low proficiency group, using the previous definition of outlier.

Our results from this related experiment are summarized as follows. For the low proficiency group, we regressed a slope of .27. The medium proficiency group had a slope of .29, and for the high proficiency group, the slope was computed to be .33. While the low and medium proficiency divisions performed essentially the same on this set of words, there was a decrease in the degree of transfer for the high proficiency division under this metric. We once again used the same bootstrapping procedure in order to determine whether any statistically reliable differences existed. For this limited set of skills, these slopes were not found to be reliably different. This result does not support the previous hypothesis that high proficiency students perform better. However, there exist a number of factors that might affect this, given the group of words considered. For instance, this set of words might be a set of 'average' words, in that they lack defining characteristics that would cause higher proficiency students to catch on or learn them faster (or equivalently have challenging factors that would hinder low proficiency students).

3.3 Predicting What Skills Demonstrate Transfer

In the prior experiments, we have demonstrated that there is a noticeable difference between some skills in terms of how much transfer occurs. However, one question we might ask ourselves is "What, if any, characteristics of words render them more likely to demonstrate transfer at a word root level?" Compared to other recent domain applications of LFA (see [Cen et al., 2006] and [Rafferty and Yudelson, 2007]), the use of a reading domain has another drastic difference- a much larger set of skills to be considered. Because of this, we can consider addressing the problem of predicting whether a particular pair of skills will be improved by being merged. This problem is equivalent to the problem of predicting what skills demonstrate transfer, by our earlier reasoning.

Up until now, attempting such classification with LFA made little sense. Even the Geometry Tutor dataset, perhaps the most recent application of LFA, had data which fit into a set of 15 skills. In order to approach the task of classification for reading, however, we must determine what features are reasonable for consideration.

We first make a distinction between two types of features pertaining to a potential merge. The first type to be considered is a set of features pertaining to the words themselves, which is domain-dependent. We chose to consider the following five features: the length of the root of the words involved, whether the root itself was one of the two words involved in the merge, whether the word root began with a vowel, the suffix of the word (which fell mainly into three buckets, those words ending in ‘s’, ‘ed’, or ‘ing’), and finally the length of the longest affix involved in the merge, where affix is defined to be the part of the word after the root itself. For instance, if we are examining a merge of the words ‘swims’ and ‘swimming’, the length of the longest affix would be 4 (since the root is ‘swim’). Our goal in performing this classification is to determine a set of characteristics of words that demonstrate transfer. This would have the potential to uncover new insights about students’ mental representations.

We trained a logistic regression model based on the results obtained on the aggregate data. Note that for all of the following tables relating to classification, a positive coefficient value represents an increased likelihood to merge as the value of that covariate increases.

Let us consider Table 3.2 and Table 3.3, containing these domain-dependent word features. Note that the absence of the '-ing' column is due to the intrinsic collinearity of the three covariates involving word suffixes. Since our dataset had few merged results with other suffixes, these three classes had to represent the entire population. We can first note that, with the exception of the statistical significance when the root is one of the words, the difference between the AIC and the BIC results is largely superficial.

There are three features that seem to have some reliable effect on the demonstration of transfer. First, the length of the affix has a negative effect on both model scores. This intuitively makes sense, since the word root may be reasonably seen as a less significant part of the word, as the word itself increases in length. The other two features are the two suffixes '-ed' and '-ing'. We note that the effect of both seems to be against transfer, which initially seems odd. One potential reason for this is that words with these suffixes sometimes will change in structure (e.g. 'study' becomes 'studies' or 'studied'). This change may make these suffixes more difficult than one like '-ing', where most of its applications require only appending the suffix to the end of the word (as per the example, 'studying').

The second type of features to consider is based upon our model fit prior to performing any merge, which is domain-independent. We described our model approximation earlier, and this model led up to the following set of three features: the absolute value of the ratio between the difficulty levels for each of the two skills (henceforth referred to as the 'Absolute Difficulty Ratio'), the ratio between the learning rates for the skills (the Learning Rate Ratio, set to 0 in the event of a negative learning rate), and thirdly a parameter for the 'student smarts', which was normalized in order to standardize its values as a covariate. Table 3.4 and Table 3.5 represent the AIC and BIC results for the model features. Once

Table 3.2: Word Feature Classification on Aggregate Data, AIC

	Word Features (AIC)						
Student Data	Root Length	One Skill is Root	Affix Length	Starts with Vowel	Suffix: -ed	Suffix: -s; -es	Suffix: -ing
Aggregate	.113 (p=.31)	-.660 (p=.01)	-.502 (p<.01)	.340 (p=.38)	-1.024 (p=.01)	-.903 (p=.03)	- -

Table 3.3: Word Feature Classification on Aggregate Data, BIC

	Word Features (BIC)						
Student Data	Root Length	One Skill is Root	Affix Length	Starts with Vowel	Suffix: -ed	Suffix: -s; -es	Suffix: -ing
Aggregate	-.132 (p=.27)	-.456 (p=.11)	-.383 (p=.01)	.266 (p=.53)	-.843 (p=.07)	-1.262 (p<.01)	- -

Table 3.4: Model Feature Classification on Aggregate Data, AIC

	Model Feature Coefficients (AIC)		
Student Data	Absolute Difficulty Ratio	Learning Rate Ratio	Normalized Student Smarts
Aggregate	-1.217 (p<.01)	.248 (p=.55)	-.020 (p=.87)

Table 3.5: Model Feature Classification on Aggregate Data, BIC

	Model Feature Coefficients (BIC)		
Student Data	Absolute Difficulty Ratio	Learning Rate Ratio	Normalized Student Smarts
Aggregate	-.617 (p<.01)	-.624 (p=.14)	.071 (p=.55)

again, a positive coefficient value represents an increased likelihood to merge as the value of that covariate increases. Similarly, a negative coefficient implies a decreased likelihood of a merge (and hence a decreased likelihood that merges with high values for that covariate will demonstrate transfer).

There are several observations we can make from these tables, concerning domain-independent features. First, the only covariate that seems to have any statistically reliable impact was the Absolute Difficulty Ratio. This result implies that as the variance in difficulty of two words with the same root increases, the less likely they will demonstrate transfer.

Overall, these attempts at classification resulted in reasonable preliminary results.

Aside from developing a novel technique to predict transfer, we have additionally documented several reliable effects discovered by this method. From these observations, we believe that there is a great deal of potential in this approach to evaluating skills within student models. Perhaps future insights will lead to more meaningful results or alternate approaches to this problem.

Chapter 4

Contributions, Future Work, and Conclusions

4.1 Contributions

Within the framework of LFA, we have made four distinct contributions. First, from a technical standpoint, we have provided advancements to the methodology behind the model. Instead of the original full statistical model, we have introduced a fast transition-based approximation. By doing so, we have rendered LFA tractable on datasets which are orders of magnitude larger than those originally used [Cen et al., 2006].

Second, we have introduced the framework of LFA to a new domain: student modeling in reading. Our new technical contributions have allowed us to demonstrate that the use of a word root model of representation can improve our representation of student knowledge, using both AIC and BIC as model scoring metrics. Despite this concurring result between the metrics, we also demonstrate that the different model complexity penalties in these metrics lead to differing degrees of transfer. This contribution not only demonstrates the utility of LFA as a modeling technique to the reading community, but it also serves the purpose of providing evidence that LFA can be used as a cognitive modeling technique in a domain independent manner.

Additionally, by examining our results along a new dimension of student proficiency, 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. Not only does this third contribution provide a technique for extracting trends by disaggregating data through LFA, but it more importantly implies that this disaggregation is nontrivial.

Lastly, we have introduced a new approach to examining skills in the context of an LFA model. Our analysis concerning prediction skills which demonstrate transfer is a novel approach in this context. Our analysis demonstrates both domain dependent and domain independent methods of performing such a prediction. We have identified several domain dependent and independent traits which have a reliable effect when predicting transfer at the word root level. It seems likely that this approach, as well as these results, hold great promise for the potential of LFA in yet another dimension.

4.2 Future Work

While this research represents a step forward both in terms of the LFA framework itself as well as its potential results, there remain a number of unresolved points. First, from a more technical standpoint, a great deal of optimization could be done to an implementation in order to make LFA into a more efficient modeling approach. For instance, we could 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.

Additionally, while we have contrasted the potential for the original model of Learning Factors Analysis and our smaller approximation, it seems reasonable that some middle road might yet be found. Perhaps a hybrid model of computation that would sometimes take advantage of the representational power of the full original model, and other times use our approximation's speed to best perform a search among potential cognitive models. For our domain instantiation, we were faced with a relatively simple search to perform. Such a hybrid model might be more useful or perhaps necessary in order to represent a more complex domain, or one in which more operators are allowed.

Our attempts to design methods of classifying and predicting operations are also a prime source for potential future work. As our results indicate, we failed at the task of separating words that merge from those that do not given our classification techniques. It is possible that focusing on a dimension of the problem such as what features to use or other methods of classification might yield more fruitful results. Such an advance would have an immediate impact on both the domain in which it was demonstrated, as well as to the cognitive modeling and AIED communities.

4.3 Conclusions

Implications of our results to relevant communities such as Intelligent Tutoring Systems, Cognitive Modeling, and Artificial Intelligence in Education abound. LFA itself is designed to be a domain-independent automated approach to finding an optimal cognitive model. Our results show that we indeed find an improved model. More importantly, though, our results imply that we have no reason to believe one of the basic tenets of tutoring systems: using a canonical model of student cognition is the best way to teach. Our results specifically imply the potential for different domain models across student proficiency. For example, we have seen that high proficiency students have a more compact domain representation than low or medium proficiency students, when considering a word root level of representation in reading.

Additionally, we have introduced a new manner of examining student cognitive models within the LFA framework. In the context of student proficiency, we have developed an approach to examining differences in cognitive skills across disaggregated student groups. Our contrast of the AIC and BIC scoring metrics illustrate how varying conclusions can be made about even an extensive dataset, by making only slightly different scoring assumptions. Both metrics come to the decision that transfer is observable at a word root level. However, their differing penalties for model complexity yield significant differences in the degree of transfer that is observed.

Next, despite failing to identify any significant results with the test gains divisions at any dimension, our other preliminary results seem to indicate that these differences are not trivial, as we previously noted upon a positive correlation between the merging of words in our model and the proficiencies of the corresponding student divisions. This result is itself surprising, given that it seems intuitive that high proficiency students would also have much overlap with students with high performance during the school year.

The approaches and results represented in this paper represent a significant step forward both in terms of developing the Learning Factors Analysis framework specifically, as well as furthering the causes of any research community involved with the task of creating, evaluating, or working with cognitive models. This preliminary work presents beneficial potential for this framework, and has offered a number of dimensions by which it can both be utilized and further developed.

Bibliography

Hirotsugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

S. Banerjee, 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.

J. Beck. Using learning decomposition to analyze student fluency development. In *ITS 2006 Educational Data Mining Workshop*, Jhongli, Taiwan, 2006.

Hao Cen, Kenneth Koedinger, and Brian Junker. Automating cognitive model improvement by a* search and logistic regression. *Proceedings of AAAI 2005 Educational Data Mining Workshop*, 2005.

Hao Cen, Kenneth Koedinger, and Brian Junker. Learning factors analysis - a general method for cognitive model evaluation and improvement. *Proceedings of Intelligent Tutoring Systems*, 2006.

Ethan A. Croteau, Neil T. Heffernan, and Kenneth 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*, volume 3220, pages 240–250. Springer Berlin, 2004.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

Neil T. Heffernan and Kenneth 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*, pages 484–489. Lawrence Erlbaum Associates, Inc, Mahweh, NJ, 1998.

R Ihaka and R Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

K Koedinger and J.R. Anderson. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8:30–43, 1997.

P Komarek. *Logistic Regression for Data Mining and High-Dimensional Classification*. PhD thesis, Carnegie Mellon University, 2004.

P Komarek and A Moore. Fast robust logistic regression for large sparse datasets with binary outputs. In *Artificial Intelligence in Statistics*, 2003.

P Komarek and A Moore. Making logistic regression a core data mining tool. Technical report, Carnegie Mellon University, 2005.

J. Meynard. Coefficients of determination for multiple logistic regression analysis. *American Statistician*, 54:17–24, 2000.

J. Mostow and G Aist. Evaluating tutors that listen: An overview of project listen. In F. Forbus and P. Feltovich, editors, *Smart Machines in Education*, pages 169–234. MIT/AAAI Press, Menlo Park, CA, 2001.

J. Mostow, 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*, Pittsburgh, PA, 2002.

A. Newell and H.A. Simon. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

Anna N. Rafferty and Michael Yudelson. Applying learning factors analysis to build stereotypic student models. In *AIED*, Marina Del Rey, 2007.

J.R. Shewchuck. An introduction to the conjugate gradient method without the agonizing pain. Technical Report CS-94-125, Carnegie Mellon University, 1994.

M. K. Singley and J.R. Anderson. *Transfer of Cognitive Skill*. Erlbaum, Hillsdale, NJ, 1989.

C.J. van Rijsbergen, S.E. Robertson, and M.F. Porter. New models in probabilistic information retrieval. Technical Report 5587, British Library, 1980.

R.W. Woodcock. *Woodcock Reading Mastery Tests- Revised (WRMT-R/NU)*. American Guidance Service, Circle Pines, Minnesota, 1998.