Refining Prerequisite Skill Structure Graphs Using Randomized Controlled Trials

by
Seth Akonor Adjei

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy
in

Computer Science

_____
April 2018

APPROVED:

_____
Professor Neil T. Heffernan, Thesis Advisor

_____
Professor Joseph E. Beck, Committee Member

_____
Professor Erin Ottmar, Committee Member

_____
Professor Jacob Whitehill, Committee Member

_____
Professor Ryan Baker, External Committee Member

Abstract

Prerequisite skill structure graphs represent the relationships between knowledge components. Prerequisite structure graphs also propose the order in which students in a given curriculum need to be taught specific knowledge components in order to assist them build on previous knowledge and improve achievement in those subject domains. The importance of accurate prerequisite skill structure graphs can therefore not be overemphasized. In view of this, many approaches have been employed by domain experts to design and implement these prerequisite structures. A number of data mining techniques have also been proposed to infer these knowledge structures from learner performance data. These methods have achieved varied degrees of success. Moreover, to the best of our knowledge, none of the methods have employed extensive randomized controlled trials to learn about prerequisite skill relationships among skills. In this dissertation, we motivate the need for using randomized controlled trials to refine prerequisite skill structure graphs.

Additionally, we present PLACEments, an adaptive testing system that uses a prerequisite skill structure graph to identify gaps in students' knowledge. Students with identified gaps are assisted with more practice assignments to ensure that the gaps are closed. PLACEments additionally allows for randomized controlled experiments to be performed on the underlying prerequisite skill structure graph for the purpose of refining the structure. We present some of the different experiment categories which are possible in PLACEments and report the results of one of these experiment categories. The ultimate goal is to inform domain experts and curriculum designers as they create policies that govern the sequencing and pacing of contents in learning domains whose content lend themselves to sequencing. By extension students and teachers who apply these policies benefit from the findings of these experiments.

Acknowledgement

I will like to take this opportunity to thank my dear wife, Esther, for her continued support and assistance through the years of hard work that has culminated in the work presented in this dissertation. I also appreciate our lovely children Bryan and Kayla for their interest in my work and for the various and sometimes subtle ways in which they have provided encouragement to me. I deeply appreciate my parents Samuel and Sarah Adjei for their love and belief in my ability to do whatever I set my heart to doing. It saddens me that my dad did not live to see the end of this journey he helped me start a few years ago.

This dissertation and years of work at WPI would not have been possible without the guidance and support of my advisor, Prof. Neil Heffernan. He, figuratively, held my hand through the challenging and yet exciting experience of graduate school education. I have really loved my time at WPI and working in Neil's lab, which is why I cannot but thank him for his support. Cristina, Prof Heffernan's wife, has been equally helpful, giving me pointers and reviewing some of the work that finally resulted in the development and release of PLACEments, the adaptive testing system within which I did most of the work reported in this dissertation. Cristina assisted in recruiting teachers for research related surveys as well as beta testing PLACEments prior to its production release.

I will like to acknowledge the help I received from colleagues like Andrew Burnett, David Magid, Chris Donnelly, Dr. Douglas Selent, Korinn Ostrow and Anthony Botelho for the various forms of support they provided. They were very helpful as peer reviewers of publications and code and, as collaborators on some of the work presented here. Anthony was especially helpful as a collaborator and, in editing my work and making helpful suggestions for improvement.

The ultimate thanks go to *Jehovah God* for giving me the life and time to carry out this interesting work, despite the many responsibilities I had to shoulder.

Contents

Table of Figures

List of Tables

1 Introduction

Prerequisite skill structure graphs have been developed by domain experts over many years, specifying the scope and sequence of knowledge components. They have been represented in several forms including learning trajectories [24, 25, 85] as well as learning maps and pacing guides [98]. Regardless of the overall representation, each knowledge component represents a given topic or skill in the knowledge domain. In considering middle school math for instance, "Addition of Fractions" and "Multiplication of Mixed Numbers" are two examples of knowledge components that may exist within the graph. These, in addition to other such knowledge components, represent math skills that students need to know at the given grade level defined by the domain experts. The prerequisite skill structure graphs further describe the order in which students in a given curriculum should be taught specific knowledge components in order to effectively build on previous knowledge and improve achievement in those subject domains. A pair of skills are said to have a prerequisite skill relationship between them if one of the skills is a prerequisite to the other (commonly referred to as the post-requisite skill); these relationships could be causal in nature, particularly if these relationships are strong. [73] In other words, when two skills have a strong prerequisite skill relationship between them, it is implied that knowledge of the prerequisite skill causes faster learning of the post-requisite skill. The importance of accurate prerequisite skill structure graphs can therefore not be overemphasized. Learners' performance in standardized tests can be attributable to, among other causes, the effectiveness of instruction that learners receive prior to these tests. Specifically, the order in which students progress through content is very important to their success [75], as measured by performance in standardized tests and preparation for future employment.

In view of the importance of the accuracy of prerequisite skill structures, also referred to as skill topologies, many approaches have been employed by domain experts to design and implement these structures. A number of data mining techniques have been used to infer these knowledge structures from learner performance data. The evolution of the study of prerequisite skill structures using data mining was largely influenced by the inference of the Q-Matrix, a mapping of items to knowledge components. [86] Several of the data mining techniques that have been employed afterwards use the Q-matrix representation of the mapping between items and knowledge components [12, 86]. Additionally, Deep Learning [67, 100], Learning Factors Analysis [19], and

Bayesian Networks and Association Rule Mining [21] have all been employed to refine existing skill topologies and also to infer new skill topologies.

While the above-mentioned approaches have arguably chalked certain degrees of success, to the best of the author's knowledge, none of them have been used to make causal claims about the prerequisite relationships among the inferred skills. Randomized controlled trials, sometimes referred to as the gold standard of research [47], have been noted to be the most effective research method to identify causal relationships between constructs in real life. [97] In view of the apparent failure of the aforementioned methods to make causal claims about the skill topologies, this dissertation presents motivation for the use of randomized controlled trials for inferring prerequisite skill structures from learner performance data. This motivation is preceded by a presentation of several techniques that this author has used to infer prerequisite skill structure graphs from learner performance data.

In this dissertation, an adaptive assessment and remediation system that uses a prerequisite skill structure graph to identify gaps in students' knowledge is presented. The system additionally assists students with identified knowledge gaps by assigning them more practice assignments to ensure that the gaps are filled. The relevance of this system to this work is exemplified through the feature that allows randomized controlled experiments to be run on the underlying prerequisite skill structure graph. The results of the experiments allow for causal claims to be made about the relationships among knowledge components. Additionally, a framework for inferring and refining prerequisite skill structure graphs is presented, with the sole aim of assisting domain experts who design these knowledge structures, and students who benefit from well-designed structures.

This thesis is organized into thirteen (13) chapters. In each chapter, related work is presented separately. In view of this, no separate chapter is dedicated to related work or literature review, however chapter 2 presents a brief description of skills and prerequisite skill structures as used in the context of this dissertation. Chapters 3 and 4 present the results of the application of a Bayesian Network-based combinatorial search algorithm to refine an existing prerequisite structure graph. These chapters show the factors that contribute to the effectiveness of such a combinatorial search method. Chapter 5 reports the findings of an investigation in which Deep Knowledge Tracing [67] is applied to infer skill topologies from learner performance data. Chapter 6 illustrates how a simple

task of predicting students' performance in post-requisite skills using learning and performance information on the prerequisite skill was used to infer prerequisite skill relationships among skills. Chapters 7 and 8 describe two different methods that this author employed to infer prerequisite skill relationships between skills. A correlation-based approach for inferring prerequisite skill links is presented in chapter 9. The results of a randomized controlled experiment in which the impact of the order of learning tasks on student assignment completion rates and subsequence performance are reported in chapter 10. Chapters 3 to 10 act to form a foundation for the primary focus of this dissertation, culminating in Chapters 11 and 12, which describe PLACEments, the adaptive learning system for diagnosing and remedying gaps in student prerequisite knowledge, and the randomized controlled experiments carried out within this infrastructure to infer relationships between skill links. Chapter 13 presents concluding remarks in regard to the work presented in this dissertation, detailing possible future work that is available through this system.

2 Background

This chapter presents a brief overview of the field of domain knowledge representations. It presents working definitions of skill, prerequisite skill, and prerequisite skill graph.

2.1 Skill / Knowledge Component

A knowledge component has been defined as "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks." [49] As noted in [49], this definition encompasses other terms (such as production rule [8], schema [46], misconception [23], skill, and concept) which define pieces of cognitive knowledge that students must learn or express either explicitly or implicitly. In the context of this dissertation, a skill or knowledge component is defined as a concept or set of concepts that students in a given cognitive domain are expected to be taught at a given level of their education. Take middle school Algebra as an example cognitive domain. Middle school Algebra can be broken down into smaller concepts (like Addition of Fractions, Multiplication of Mixed Numbers, etc.) each of which students are expected to be taught at a certain grade level in the course of middle school training. Each of these concepts is defined as a skill. In this write-up, skills and knowledge components are used interchangeably. They represent the knowledge required by a learner to be able to solve specific types of problems. A typical example problem for the "Addition of Fractions" knowledge component is:

$$\frac{2}{5} + \frac{5}{9}$$

For this example, any student able to correctly answer this question and all other possible questions of this format is said to have acquired the "Addition of Fractions" knowledge component. This is a single component of a larger group of components. Groups of well-defined skills form a cognitive domain. Others have defined a domain as a group of different problem types (or knowledge components). [34] Each individual learner has a well-defined subset of problem types (or knowledge components) that they can comfortably solve under normal circumstances (i.e. not under any emotional or physical pressure). This set of knowledge components is referred to as the student's knowledge state. For every given set of knowledge components, there is a large number of

possible knowledge states that can be exhibited across students through the acquisition of varying combinations of knowledge components.

2.2 Prerequisite Skill / Skill Graphs

Webster's International Dictionary defines learning as "*the activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something*." This implies that learning is a process, and that as a student progresses through the learning process, the number of knowledge components in a learner's knowledge state is expected to increase. This process involves building on an already-existing knowledge state. For many concepts, there is a set of knowledge components that a learner is expected to know in order for the learning of the new unknown knowledge components to occur in a less arduous manner.

Two skills are said to be in a prerequisite relationship if learners are required to know one of the skills in order to more easily learn the second skill and subsequently correctly answer questions of the second skill. Figure 2-1 below shows a sample prerequisite skill link between two skills: *Addition of Fractions* and *Multiplication of Fractions*. The arrow points from the prerequisite skill to the second skill in the pair. This second skill is also referred to as the post-requisite skill of the first skill. As depicted in the diagram, a student is expected to know how to correctly answer questions/problems that test their ability to add fractions in order to be able to correctly respond to questions relating to multiplication of fractions.

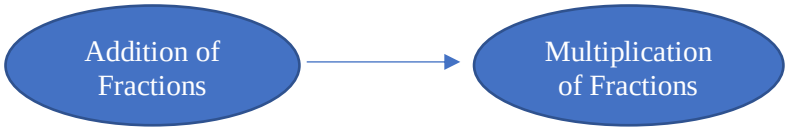


Figure 2-1 An Example Prerequisite skill link

Some skills have multiple prerequisite skills. For those cases, it is implied that knowledge of all the prerequisite skills is necessary and required for students to be able to easily learn and correctly respond to problems of the post-requisite skills. This relationship is referred to, in most literature, as a conjunctive relationship. [29] Disjunctive relationships are those in which not all the prerequisites are required for success in the post-requisite skills. In the context of this dissertation and for cases where a skill has more than a single

prerequisite skill, we assume a conjunctive relationship and do not claim to make any statements about the disjunctive links.

A prerequisite skill graph represents a collection of all the skills in a given cognitive domain as well as the relationships between the constituent knowledge components. It is usually depicted as a directed acyclic graph in which the nodes represent the knowledge components and the arrows indicate the direction of prerequisite relationships. In the context of this dissertation, we avoid cyclic graphs (graphs that contain cycles). We hypothesize that graphs that have cycles cause confusion regarding which skills to blame when learners fail to demonstrate learning for any selected knowledge component in the cycle, and hence beyond the scope of this work. Furthermore, the presence of a cycle is indicative that no prerequisite relationship exists among the skills in the cycle since this will be arbitrary as to the ordering of skills within a cycle. Figure 2-2 depicts an example of these domain-expert-designed graphs. In this particular example, the rectangles represent knowledge components, the links between them indicate the direction of the prerequisite relationship, and the colors are meant to show the different grade levels at which students are expected to be taught different knowledge components. Figure 2-3 shows another example of the prerequisite skill structure. In this particular example, skills are represented as hexagons, and the attachments to other hexagons show the relationships between them.

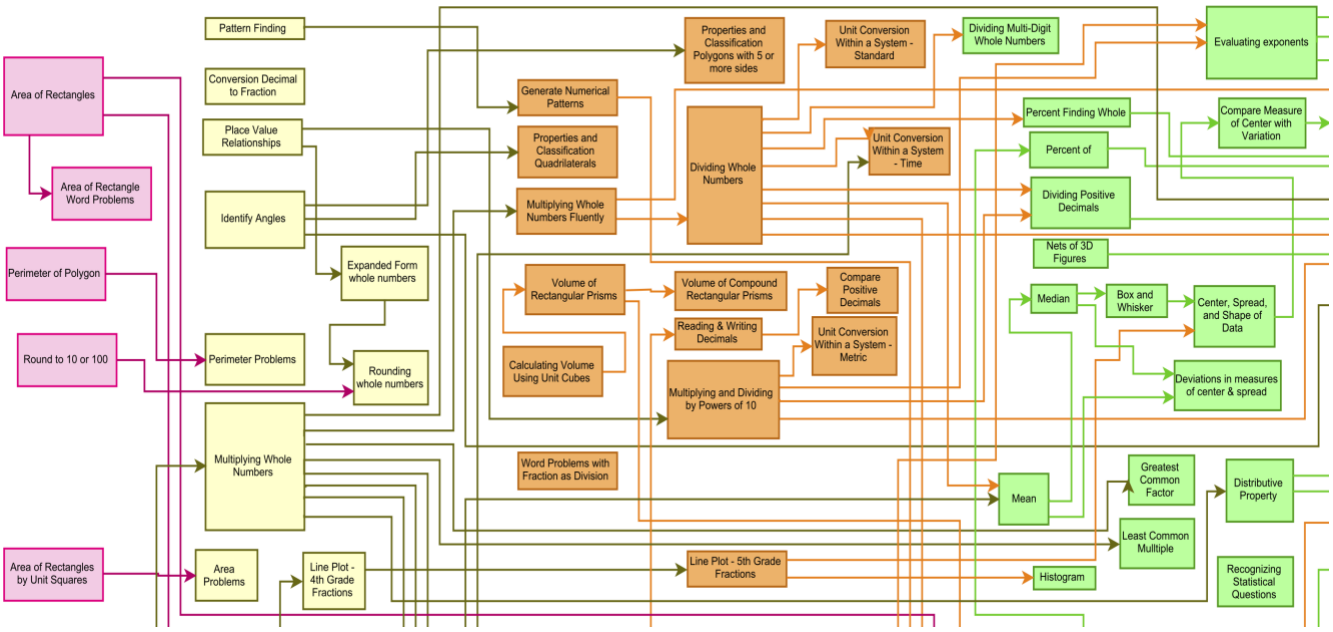


Figure 2-2 An example prerequisite skill graph

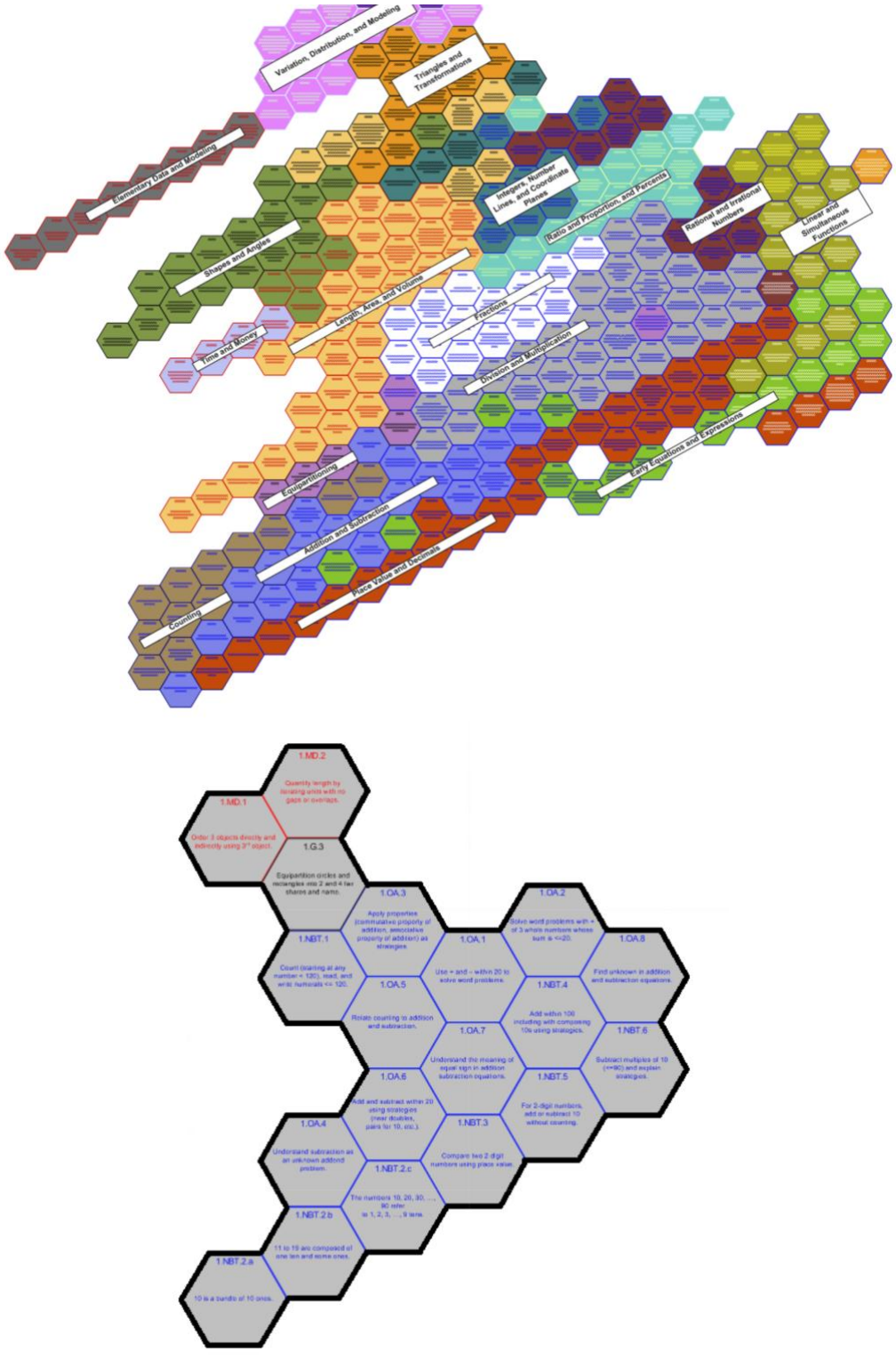


Figure 2-3 Another example graph sourced from [turnonccmath]

The chapters that follow describe several attempts that have been made to refine structures of this nature, and the degrees of success that this author and other authors have reported in this area of research with the ultimate goal of supporting teachers, and curriculum developers to improve the educational experience of students.

3 Refining Learning Maps with Data Fitting Techniques: Searching for Better Fitting Learning Maps

Learning related sciences need quantitative methods for comparing alternative theories of what students are learning. An example of models that represent what students are expected to learn in a given domain is the learning map. A learning map, sometimes referred to as a concept map, is the graphical representation of the skills and knowledge components that students are expected to learn in a subject area (such as math) and the relationships between these skills/concepts. [15, 81] This chapter presents an investigating of the accuracy of a learning map and its utility to predict student responses. Our data included a learning map detailing a hierarchical prerequisite skill graph and student responses to questions developed specifically to assess the concepts and skills represented in the map. Each question aligned to one skill in the map, and each skill had one or more prerequisite skills. Our research goal was to seek improvements to the knowledge representation in the map using an iterative process. We applied a greedy iterative search algorithm to simplify the learning map by merging nodes together. Each successive merge resulted in a model with one skill less than the previous model. We share the results of the revised model, its reliability and reproducibility, and discuss the face validity of the most significant merges.

A version of this chapter is published at the following venue:
Adjei, S. A., Selent, S., Pardos, Z., Broaddus, A., Heffernan N. & Kingston, N. (2014) Refining Learning Maps with Data Fitting Techniques: Searching for better fitting learning maps. In John Stamper et al. (Eds) *Proceedings of the 7th International Conference on Educational Data Mining*.pp413-414

3.1 Introduction

Cognitive models are used to represent how one's knowledge may be organized [41]. As such, they contain descriptions of component pieces of knowledge and connections among the components to indicate how understanding develops in a specified domain [41]. Different authors have described various cognitive models, including learning maps [68], learning trajectories [24], and learning hierarchies [40]. Learning maps use linear sequences of learning goals and are useful for instructional planning [68]. A learning trajectory includes a learning goal, a developmental progression defining the levels of thinking students pass through as they work toward the defined goal, and a set of learning activities or experiences that assist students in reaching the defined

goal [24]. As their name implies, learning hierarchies model prerequisite knowledge components in hierarchies, allowing multiple pathways to extend from one prerequisite skill to multiple learning goals [40].

The learning map extends the notion of a learning hierarchy by representing domain knowledge as a network of component skills and connections, allowing for multiple paths from prerequisites to learning goals. While multiple paths add complexity to the cognitive model, they allow the learning map to represent the potential learning of a broad range of individuals who may experience difficulties traversing certain pathways due to disabilities or particular learning preferences. As such, the learning map provides a flexible model of learning that is consistent with recent advances in universal design for learning [15, 80].

In the present study, we examine a small section of the learning map and investigate the effects of permuting the topology of the hierarchy. Skills and concepts are represented by latent nodes in the learning map. Directed edges represent the prerequisite relationship among latent nodes and also represent the relationship between those nodes and their associated test items. We present a simple method for improving the predictive power of the learning map by combining latent nodes. We report our initial results on the fit improvement, stability of the resulting map, and interpretation of the algorithms chosen node combinations.

This work connects with literature on searching for better fitting cognitive models. Several non-hierarchical cognitive models have been developed to represent the relationship between knowledge components (KCs) in the form of prerequisite skill maps. These cognitive models have been developed to help intelligent tutors, as well as experts, determine student mastery of KCs. A number of technical approaches have been developed to evaluate cognitive models developed by domain experts. One approach is Learning Factors Analysis (LFA), developed by Cen, Koedinger and Junker [19] to help the Educational Data Mining (EDM) community evaluate different cognitive models.

There are several different methods for analyzing skills. Tatsuoka [86] introduced the rule space method for representing and determining how well students understood the underlying skills (or rules as the authors call it) for test items. Additionally, the method is used to identify any erroneous classification or misconceptions of students in responding to test items. Barnes [12] utilized the Q-matrix method from Tatsuoka's rule space method to organize combinations of skills into distinct latent classes and assign students to latent classes based on level of

mastery. Additive Factor Models (AFM) also utilize the Q-matrix but with a multiple logistic regression model which predicts student performance based on a number of factors, primarily the number of opportunities a student has to demonstrate a particular skill. Cen reported in [19] that AFM did not accurately predict items involving conjunctive skills and hence introduced the Conjunctive Factor Model (CFM) to improve predictions in this area. In addition to latent skill cognitive models, item to item knowledge structures have also been learned from empirical data using Bayesian Network structure learning and partial order knowledge structures [30].

Our approach to simple merging of skills was inspired by Learning Factors Analysis [19], which uses a combinatorial search to determine which model best fits student data. The combinatorial search consists of three different types of operations: splitting, merging or adding existing KCs. Splits occur when a knowledge component is determined to be composed of more than one skill, and hence splits into multiple skills. One or more skills are merged if they are determined to be inseparable skills, given student data. The add operation involves the inclusion of a completely new skill to the original map [18].

Other researchers have tried to extend LFA to other subject domains. Leszczenski and Beck introduced a scalable application of the LFA framework in the context of reading knowledge transfer [52]. The problem with this approach is that the search was unstable and could give different results each time the search was run. Instead of determining a student model given an initial human generated model, Li, Cohen, Noboru, and Koedinger proposed a method for automatically generating the KCs from student responses to individual items. [53] Although their method resulted in the best fit among the other candidates, it may not generalize for models with less coarse-grained KCs. Other models have focused on the determination of a student's knowledge of certain skills. Logistic regression has been used to trace multiple sub-skills of a given skill [99]. Pavlik, Cen, and Koedinger proposed a method for automatically deriving a cognitive model by generating a Q-matrix, which provides a representation of the KCs required for each test item. [63]

In this work, we follow the process described by Cen, Koedinger and Junker [19]. This technique can be used to analyze hypothesized learning maps and consider whether small improvements to the model result in a better fit to the data. In this method two different approaches were studied to determine the best skill map from an initial graph. Cen, Koedinger, and Junker suggested three types of operations, i.e., merges, splits, and adds. [19]

However, in this study, we used only merge operations given the already highly granular quality of our initial, subject matter expert derived learning map.

3.2 **Initial Learning Map**

This study examined a section of the learning map containing 15 concepts and skills related to understanding integers. The map was developed using mathematics educational literature describing how students learn to understand and operate with integers. The set of integers includes the whole numbers and their opposites, presenting many students their first exposure to negative numbers (Van de Walle, Bay-Williams, Karp, & Lovin, 2014). Although many students have prior knowledge of negative values within contexts such as debt or temperatures below freezing, they often struggle when first learning to work with negative numbers. Proficiency with integers includes understanding opposite numbers, comparing integers, representing integers on number lines and graphs, and using integers in real world problem contexts. The learning map shown in Figure 3-1 illustrates the component concepts and skills that comprise such understanding. This map suggests that students should learn to identify opposite numbers (M-1104) and integers (M-1289) in preparation for comparing and ordering integers (M-1133, M-1135, M-1140) as well as representing integers on number lines (M-1118, M-1120, M-1108, M-1126) and coordinate planes (M-1122, M-1124). Because integers challenge the initial counting strategies students learned for positive numbers, it is beneficial for students to work with integers in real-world contexts (M-1106, M-1105, M-1127, M-1128) [90]

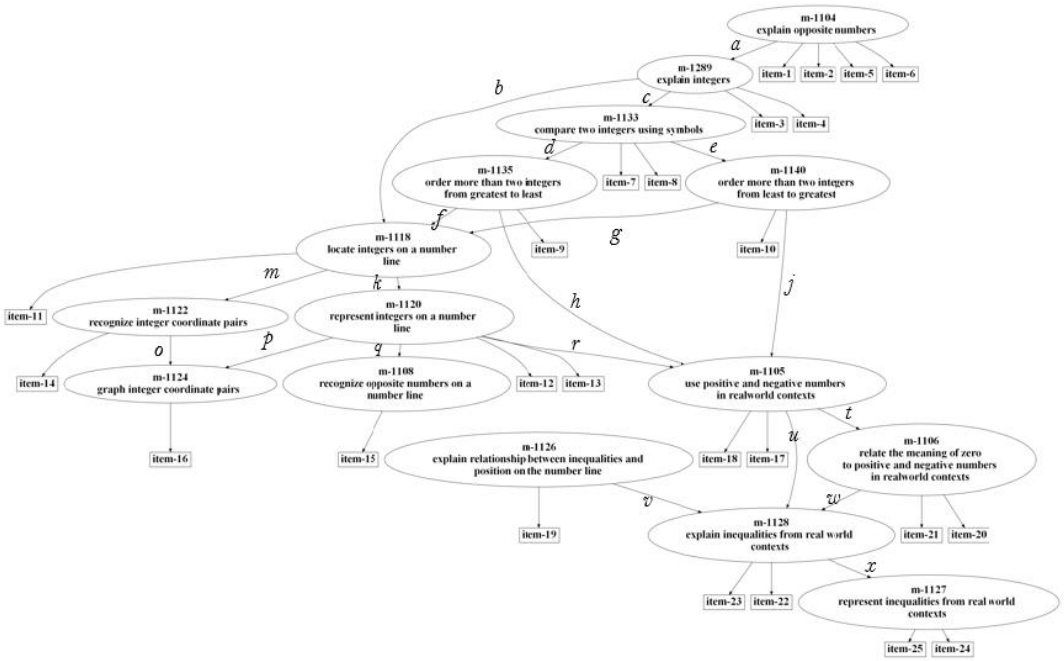


Figure 3-1 The initial learning map that researchers created. Each ellipse represents a "skill" and each rectangle represents a test item. For easy reference, the links are labeled. The labels do not have any specific meaning

The data for this study was gathered from student responses to 25 test items aligned to the 15 skills shown in the learning map in Figure 3-1. All of the test items were multiple choice questions, with four answer options per question. Each skill was assessed by one or more items. As part of the test development process, subject matter experts confirmed the alignment of each item to its associated skill, meaning that the item was judged by experts to evoke the intended skill. Therefore, when a student answered a test item correctly, we assumed in this study that the student had mastered the skill associated with that test item. Furthermore, due to the hierarchical structure of the learning map, items associated with skills lower in the learning map were assumed to be more difficult, i.e., require more skills, than items associated with skills higher in the learning map.

In addition to the graph, we utilized a data-set containing the responses of 2,846 students answering the same sequence of 25 items in the learning map. All the students were chosen from middle schools in a mid-western state from grades 6 (8%), 7 (49%), 8 (39%) and 10 (4%). The students' responses were dichotomous, '1' for correct and '0' otherwise.

3.3 **Methodology**

3.3.1 Merge Operation

In all of the experiments our sole manipulation of the map was to merge latent nodes. A merge operation occurred when two skills adjacent to each other in the map were combined into one skill. Items from both skills that were merged were reattached to the new single skill. The prerequisites of the constituent skills became prerequisites of the merged skill and the same applied to the post-requisites. An example merge operation on a section of the skill map is shown below. The skill maps before and after the merge operation are shown in Figure 3-2. M-1289 and M-1133 are the skills that were merged into a single skill, named "M-1289XM-1133". Note that the names of the skill hold no meaning of their own, just as the labels of the arcs between the skills.

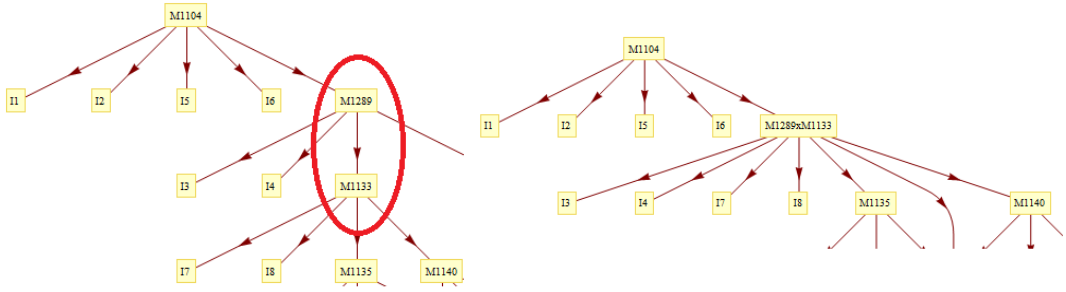


Figure 3-2 Before and after the merge of the arc between M-1289 and M-1133. Note that after the merge, all the items mapped to both M-1289 and M-1133 now are mapped to the joint skill labeled "M1289xM1133".

3.3.2 Evaluation Procedure

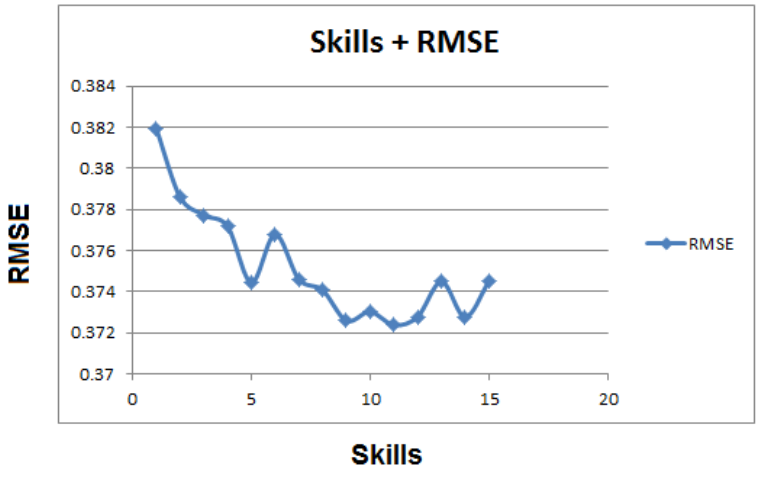
For evaluating the models, we used per student per item cross validation with 5 student folds and 3 item folds. Our student and item folds were chosen randomly for our evaluation. More details about how the cross-validation was done can be found in the technical document (1). We used the Root Mean Squared Error (RMSE) metric to evaluate the results of the experiments. RMSE is calculated by squaring the differences between each actual value and predicted value and then finding the average value of the differences. Taking the square root of the average will give the RMSE value for the model. The closer the RMSE value is to 0, the more accurate the model is (i.e. the smaller the error is in predicting the available data.)

3.4 Experiment 1: Iterative Search

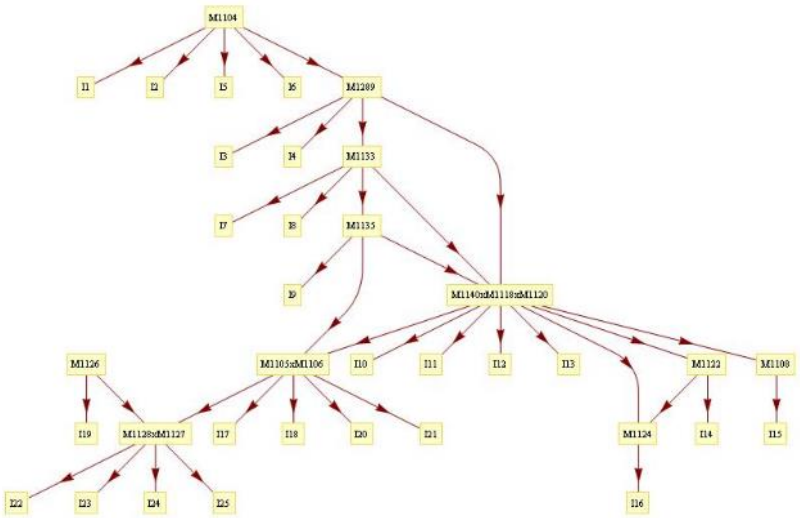
The purpose of this experiment was to take the original learning map and to create and run a search algorithm to find a better, more predictive, learning map. This experiment uses a greedy search algorithm to generate the new models. In this experiment, we started with the initial learning map shown in Figure 3-1 and created a Bayesian network to represent this map. Starting with the original map we programmatically found all possible skill pairs that could be merged. The algorithm only considered merging adjacent skills, or skills that shared an edge between them. Each possible merge was evaluated using the procedure previously described, and the best merge was chosen based on the map with the lowest cross-validated prediction error. We applied the best possible merge to the map and this resulted in a map with one less skill. The new map was used as the input to the next iteration of the algorithm. This technique was iteratively applied until all the skills were merged into a single skill. Further details of the iterative search algorithm can be found in the technical document.

3.4.1 Results and Analysis

Figure 3-3 shows a graph and an image of the prerequisite skill graph of the results from the iterative search. The search started at iteration 0, which was the initial skill map consisting of 15 skills before any merges were applied to it. The search ended at iteration 14, which is a graph consisting of just one skill with all the items attached to that one skill. The best models from each iteration are shown below. We recorded AUC, RMSE, accuracy, AIC, and BIC metrics, although we only used RMSE to choose the best models at each iteration and to guide our search. Ultimately, we chose RMSE as the deciding metric.



(a) Skill Accuracy



(b) Best Model Skill Map (11 skills at iteration 4)

Figure 3-3 a) The chart of results and b) the graph of the best skill model

The results show that the best RMSE obtained was from the 11-skill map at iteration 4 with an RMSE of 0.37238. This is slightly better at predicting students' real performance data than the original skill map (with RMSE of 0.37451). The 11-skill map has a small but significant improvement (p = 6.22624E-68) from the original skill map. The graph shown in Figure 3-3b also shows that models consisting of between 9 and 12 skills have similar RMSE values and are alternative choices for a best model depending on the level of skill granularity desired. Though the 11-skill model is significantly better than the original model (in terms of RMSE), practically these two models are the same since the difference between these RMSE values is about 0.0022. The only advantage that the 11-skill model has over the original 13-skill model is the reduction in complexity of the model.

In addition to looking at which model best predicted actual responses, we examined which skills were being merged throughout our iterative search to see if we could find any general trends. A list of the merges can be found in the technical report. The individual skills are represented by their original numbers and a merged skill is represented by the numbers of each skill concatenated with an 'x'. The numbering is in topological order, meaning that the skill highest up on the skill map was listed first for a merged skill. The first merge occurred for skills M-1128 and M-1127. Since skill M-1128 was a parent of skill M-1127, it is listed first in the combined skill name M-1128xM-1127. (See Figure 3-3b)

3.5 **Experiment 2: Stability Experiment**

In the previous experiment, every model was evaluated once and only once, which lead to the question of whether or not our results were stable. Our model evaluation used the Expectation Maximization (EM) algorithm, which is known to be affected by the starting value. For our original experiment, we chose our starting points for EM randomly and only evaluated each model once. The authors, in earlier research, found that the starting point of the EM algorithm could make a difference in the converged value. In general, the EM algorithm does converge to the correct value, but there are cases where it can converge to incorrect values or to the "opposite" value. Considering the range to be between 0-1, if the actual true value of a parameter was 0.3, EM could converge to (1 - 0.3) = 0.7 instead if the initial starting point was too far from the true value.

Our question was: if we were to run the iterative search experiment several times would we end up with the same results using different starting values for EM. Since it takes several hours just to evaluate a single model, running the entire search consisting of over 100 models to evaluate would take too long. The purpose of this experiment was to evaluate just the first iteration of the search ten times to see if the results converged to a single best graph.

For the first iteration of the algorithm there were sixteen possible merges that could happen. For each of these possible merges we evaluated the resulting model ten times. The evaluation used was the same evaluation as the iterative search experiment for which we tested stability. For each of the ten runs we set the random seed in MatLab to correspond to the run number. This gave us a different set of random numbers for each run of the 16 possible merges, where each merge got the same random seed within a run. Manually setting the random seed also meant our results for the stability experiment would be reproducible

3.5.1 Results and Analysis

After evaluating all sixteen models from the first iteration ten times we kept a count of how many times a model was the best model and how many times a model was in the top 3 best models. RMSE was used to choose the best models since it was used to determine the best model in the iterative search experiment. The results are shown in the table below in Figure 3-4.

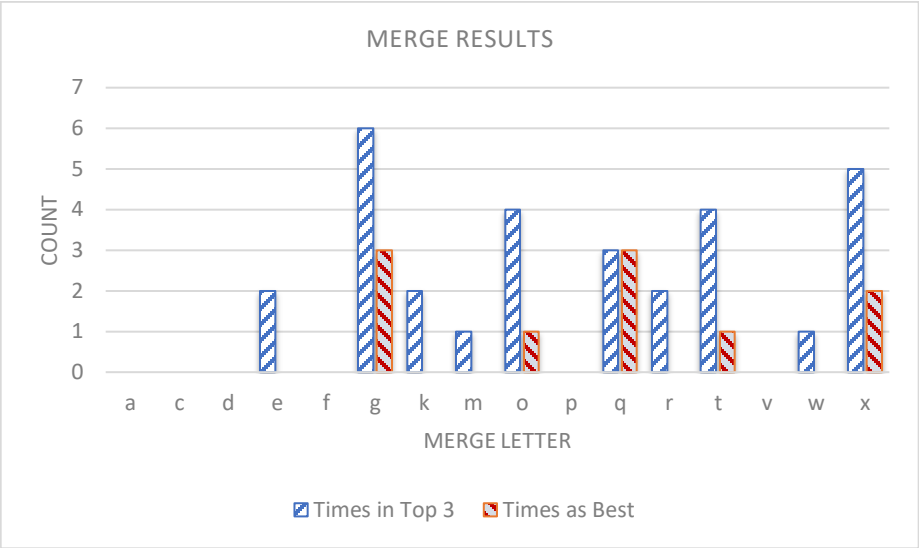


Figure 3-4 Stability of Graph

Merge ‘g’ was in the top 3 the most times (6) and was also the best model the most times (3). Merge ‘x’ and merge ‘q’ also did well. Merge ‘x’ was in the top 3, five times and was the best model two times. Merge ‘q’ was in the top 3, three times and was the best model three times. Merges ‘t’ and ‘o’ also did well. The general observation was that there was separation between good and bad merges but the best merge was not stable and did not converge.

We compared the graphs to our original iterative search experiment. In the original iterative search, the first two skills that were merged were skills M-1128 and M-1127, corresponding to merge ‘x’ in our stability experiment. The second two skills that were merged in the iterative search experiment were skills M-1140 and M-1118, corresponding to merge ‘g’ in the stability experiment. Both merges ‘x’ and ‘g’ were the best two graphs in the stability experiment. Although merge ‘g’ did slightly better in the stability experiment, the order in which the merges took place did not matter. The best model in the iterative search took place after 4 merges, which included merges ‘x’ and ‘g’. Although we could not run the stability experiment 10 times for all possible merges and merge paths, we believe that it has a decent chance to converge to the same best model, which occurred after the fourth merge in the iterative search.

3.6 **Discussion**

When analyzing each merge, we considered the skills or concepts described by the affected skills as well as the test items associated to those skills. The descriptions below discuss the three groups of skills merged in

experiment 1 and shown in the Best Model Skill Map (Figure 3-3b). The two additional pairs of skills merged in experiment 2 are also discussed. In each case, the merges point to commonalities in the skills themselves or among the test items used to assess different skills.

Merge ‘x’ affected skills M-1127 and M-1128. These skills represent “the abilities to represent inequalities from real world contexts” and “explain inequalities from real-world contexts”, respectively. The test items associated with these skills required students to read problems and identify inequality statements that matched the problems. In this case, the test items did not distinguish between two unique skills, i.e., representing a problem or explaining a problem, as was suggested by the two skills.

Merges ‘g’ and ‘k’ affected skills M-1118, M-1140 and M-1120. These skills represent the abilities to “locate integers on a number line”, “represent integers on a number line”, and “order integers from least to greatest”, respectively. The test items associated with these skills required students to select lists of correctly ordered integers or identify the correct number line graph of a particular integer. In this case, the test items did not adequately distinguish between locating and representing integers on a number line (i.e., M-1118 and M-1120) because all of the items were multiple-choice, and none provided students the opportunity to construct their own number line representations of integers. The inclusion of ordering integers from least to greatest (i.e., M-1140) with the other two skills is possibly due to the fact that using a number line is inherently, cognitively connected to ordering numbers from least to greatest.

Merge ‘t’ affected skills M-1105 and M-1106. These skills represent the abilities to “use positive and negative numbers in real-world contexts” and “relate the meaning of zero to positive and negative numbers in real-world contexts”, respectively. The test items associated with these skills required students to interpret problems involving integers and choose integer answers or verbal statements about integers. Two of the four test items included references to zero either as freezing point or sea level. In this case the items were designed to distinguish between the two skills, i.e., using integers and relating integers to zero. However, the relationship between zero and positive or negative numbers is so critical for understanding integers that, it is likely one cannot compare integers without considering their values in relation to zero.

Merge ‘q’ affected skills M-1120 and M-1108. These skills represent the abilities to “represent integers on a number line” and “recognize opposite numbers on a number line”, respectively. The test items associated with these skills required students to identify the correct number line graph of a particular integer or the opposite of a given integer. In this case, the two skills are inherently connected by the very definition of an integer as the opposite of a whole number. Consequently, it is likely that once students understand the definitions of integers and opposites and can use a number line, the act of graphing an integer is the same as graphing an opposite.

Merge ‘o’ affected skills M-1122 and M-1124. These skills represent the abilities to “recognize integer coordinate pairs” and “graph integer coordinate pairs”, respectively. The test items associated with these skills required students to identify the graph of a given integer ordered pair or to select the description of how to graph a given ordered pair on a coordinate plane. In this case, the items did not clearly distinguish between the two skills because the items associated with recognizing integer coordinate pairs included graphs. Furthermore, the skills themselves are difficult to distinguish in a practical sense because when students learn to graph integer ordered pairs, they routinely associate the numerical representation (i.e., the ordered pair) with its graphical representation (i.e., the point graphed in the coordinate plane).

An additional observation is that some of the skills tended to merge by pairing up with one and only one adjacent skill before RMSE started to decline. Before merge ‘t’, the merges were all pairwise with the exception of merge ‘m’. After merge ‘t’, the skills tended to keep merging into the same skill. The best skill map was generated after merge ‘r’, suggesting that adjacent skills tended to be similar skills and skills M-1140 and M-1120 were similar although they were not adjacent. This was a stronger relationship for several reasons. Firstly, the merges that culminated in the merger of M-1140, M-1118, and M-1120 all took place before the best skill map was reached. This indicated that those three skills give better predictive performance when represented as one skill. Secondly, this was the first and only 3-skill group to be merged in the best model before RMSE declines. Lastly the three skills took two iterations of the search algorithm to merge together because skills M-1140 and M-1120 were not adjacent skills. Despite the initial graph topology, our search decided to merge these three skills. The combination of all these factors provided strong reasoning that the three skills M-1140, M-1118, and M-1120 were not really distinct skills.

3.7 **Contributions, Conclusions and Future Work**

In this work, we provided a search algorithm to reduce the complexity of a given learning map, while improving its fit to real student data. Since merging skills increased accuracy, these results suggest that the original skill map was too fine-grained (given the number of questions per skill and the number of students who took the test.). In some cases, the test items did not adequately distinguish between the skills that were merged; hence such skills were merged. The results of algorithms like this can help the content experts who are creating skill maps and test items to either reconsider thinking of two skills as separate, or prompt them to write different test items to better distinguish between students that have mastered one of the skills but not the other skill. In this work, the team that created the learning map expected item 11 was a prerequisite for items 12 and 13, but our stability results suggested that of all the arcs, this arc was the least supported by the data (see Figure 3-4, arc "g"). In fact, due to this work, we asked an unbiased teacher who did know what our mapping was, to create a hierarchy between items 11, 12 and 13. Surprisingly, she suggested that 12 and 13 were prerequisites to item 11, suggesting that the arc should point in the exact opposite direction. This may indicate that our method may be helpful in using the data to suggest places in the skill graph that need more attention and refinement.

We can relate this work to our other work. Heffernan's ASSISTments project is a project that is attempting to track and improve students' knowledge across middle school mathematics. About a decade ago we had a learning map with over 300 skills but we now have reduced that complexity to 147 skills. Curriculum designers will correctly be thinking about the subtle ways in which problems are different from one another, which cause them to want to add skills to the skill maps to make more subtle distinctions between questions. However, if you also want to use the hierarchy to track knowledge, having more skills creates complexity, as few questions for each skill make fitting quantitative models harder.

All of the work we have done in this chapter has a very small number of questions per skill. This naturally would cause us to think that many merges would be necessary, but if we had a large number of questions, and added all those students' responses to that large number of questions, we could probably justify more complicated models.

In our experiments, we examined the effects of merging skills on an existing learning map. There are many other ways we could have used the existing map to create alternatives. For instance, Cen, Koedinger and

Junker [19] have explored ways of splitting skills or adding new skills, but all of those make more complicated models. What was not examined were the split and add operations. Possible future work could examine those operations to see if a better model can be obtained with them. Additionally, to validate our algorithm, applying it to synthetic learning maps and synthetic data could be useful to determine if our algorithm does converge to a true learning map.

End Notes

(1) The dataset, evaluation algorithm, and a technical report describing the algorithm in detail can be found at https://sites.google.com/site/assistmentsdata/kansas-project

4 Refining Learning Maps with Data Fitting Techniques: What Factors Matter?

Attempts have been made to refine cognitive models/Learning maps (skill graphs) using some data mining techniques. [16, 21, 32, 63, 92, 100] However, the factors that affect these improvements/refining processes are not so clear. In the previous chapter, we presented a method for improving these cognitive models. The purpose of this chapter is to present the factors to consider when using our initial algorithm to refine learning maps. We present a simulation study that shows how important each of the factors is for this refinement process.

4.1 Introduction

Learning maps have been used as a tool to depict the set of skills in a cognitive domain and the relationship between these skills. A number of studies have been conducted to find and represent the relationships between the skills [37, 58, 84, 86]. Tatsuoka introduced the Rule-space method for identifying skills/knowledge components in a given cognitive domain whereas [51] present another approach called the Attribute Hierarchy Method (AHM). The rule-space method (RSM) does not present the relationship of the skills/knowledge components as a hierarchy. However, AHM, which is a variation of Tatsuoka's RSM, considers the hierarchical relationship between the components [87]. Gierl used the AHM approach to make inferences of students' cognitive assessment. [41] None of these approaches dealt with methods for improving the item response models developed using the methods proposed. The Learning Factors Analysis (LFA) method [19] was introduced to deal with this problem. In that chapter three different operations for improving the predictive abilities of learning maps or cognitive models are introduced. In [1] an attempt was made to solve this problem by presenting the results of a number of experiments that showed that learning maps can be refined using just one (the merge operation) of the three possible of the LFA method. It was shown that there were significant improvements in RMSE for the best model chosen, starting off with a pre-defined learning map.

We realize that, to generalize the method for refining learning maps, there are a number of questions that still need to be answered. These include: "What are the factors that can determine when a model can be best refined?" and "Do the number of skills, the number of items per skills, the number of levels in the skill hierarchy and the number of data points have any effect in determining the best refined model?" Whilst the LFA methods use a set of factors to determine whether to merge, add or split skills to generate better models from an existing one, all the factors used are based on expert knowledge and are independent of data. In order to answer the above questions, we present a number of simulation experiments.

4.2 **Problem Statement**

The LFA model uses three operations (splits, merges and adds) to refine knowledge components. In each of the operations, learning factors were included in the model refinement process. These factors did not include the number of skills in the model, the levels in the hierarchy of skills in the model, the number of items per skill and the guess and slip parameter values for the items. We hypothesize that these factors are important in generating an optimal model from a given learning map (pre-requisite skill hierarchy). Hence, we set out in this chapter to present a series of experiments that help in determining the impact of the above-mentioned factors in refining a given learning map or cognitive model.

4.3 **Methodology**

To answer the research questions, we started off with a 3-skill graph. We inserted a fake skill at different locations of the graph and ran our evaluation code to determine when the original skill-graph is learned back and what factors determine when this occurs. We define a fake skill as a non-existent skill within the current domain or in any other domain for that matter. The intent is to determine whether our iterative method [1] can identify such a skill and eliminate it from the final refined skill graph. We examine the following factors and determine which of these factors have the most impact on using the greedy algorithm presented in the earlier paper to refine a given model: guess and slip parameter values, the number of levels in the skill graph hierarchy and the number of data points (i.e. students and items). For each randomly chosen skill graph we generate a set of simulated data, one

each for the number of student and item pairs used. We then evaluate the models using Expectation Maximization to determine the factors that have the most impact. The section presents the random graph generation, Bayesian network creation, fake skill creation and the evaluation code.

4.3.1 Random Skill Hierarchy Generation.

To generate a skill graph randomly we start by choosing a random skill hierarchy. Our algorithm to generate the skill hierarchy takes a range of skills and a graph depth as input parameters. The output of the algorithm is a valid skill hierarchy where the number of vertices is within the skill range and the number of levels is within the depth range. We order our vertices from 1 to N and use the constraint that a vertex cannot have a directed edge pointing to a smaller numbered vertex. We also enforce the constraint that a vertex cannot have any self-edges.

To generate a random graph, we choose a random number within the range of possible graphs. We then convert this number to binary form and add the correct number of leading zero's (we know the number of skills from the random number chosen). Then we simply insert the bits of the binary number into the varying spots of the matrix form of the graph in order.

The result is a directed acyclic graph with no self-edges. It will not necessarily be completely connected, that is some of the skills be stand-alone without any prerequisite relationships with other skills. The final step is to check if the graph is connected. If the graph is connected, we keep it; otherwise we discard it and repeat the generation process. This method allows us to instantly generate valid graphs. An example is shown in Table 4-1 and Figure 4-1 for a graph with three skills.

Table 4-1. Example Matrix. Matrix generated by the random number 5. A 'Y' represents that this cell is ignored because it must be a zero since a vertex cannot have directed edges pointing to vertices with larger numbers. An 'X' represents that this cell is ignored because it must be a zero since a vertex cannot have self-edges.

Vertex/Vertex	1	2	3
1	X (0)	1	0
2	Y (0)	X (0)	1
3	Y (0)	Y (0)	X (0)

Figure 4-1. Example Graph Generated

4.3.2 Create Bayesian Network.

The Bayesian network used for the analysis was generated from the skill graph selected from the previous step. To generate the items for the skills an item range is specified. A random number of items are chosen within the item range for each skill. In our experiments, we restricted our range to be a single value so all skills will have an equal number of items. We set our Bayesian network up like knowledge tracing, where every skill has one or more items and every item has a guess and slip node. [26] An item must belong to exactly one skill. The skill nodes are latent nodes since we cannot observe whether or not a student knows the skill. Each item node is an observable node, which is a '1' if the student answered the item correctly and a '0' if the student did not answer the item correctly. Both the guess and slip nodes are also latent nodes representing whether or not the student guessed or slipped on the item. A student is considered to have guessed when the student answered correctly but did not know the skill. A student is considered to have slipped when the student answered incorrectly but knew the skill. Using the previous skill graph example, we added the item, guess, and slip nodes to the graph.

The final step to create the Bayesian network is to create the conditional probability tables (CPT) for the nodes. For our experiments, if a skill has multiple prerequisites, we considered the prerequisites as conjunctive. This means that a student should find it difficult to learn a post-requisite skill if the student does not know all of the prerequisite skills. Therefore, if a student does not know one of the prerequisite skills then the student may

find it challenging to learn the post-requisite skill. If the student does know all the prerequisite skills (or there are no prerequisite skills), we pick a random probability that the student will know the post-requisite skill between 0.3 – 0.7. Our guess and slip parameters have varying probabilities since that was one of the parameters we experimented with. All the item nodes have a deterministic (0% chance or 100% chance of correctness) CPT based off of the skill, guess, and slip nodes (which or not deterministic).

4.3.3 Creation of Fake skill

We exported our Bayesian network to Matlab and used Kevin Murphy's Bayes Net Toolkit to generate simulated data, which we define as ground truth data for the graph. Once the ground truth data was generated we randomly generated "fake" skills from the original graph. A fake skill is generated by randomly choosing a real skill. Once a real skill is chosen, a random number of items is chosen from the real skill. These items are then detached from the real skill and attached to the fake skill. The fake skill is then randomly chosen to be either a parent or a child of the real skill. Figure 4-2 shows the creation of a fake skill.

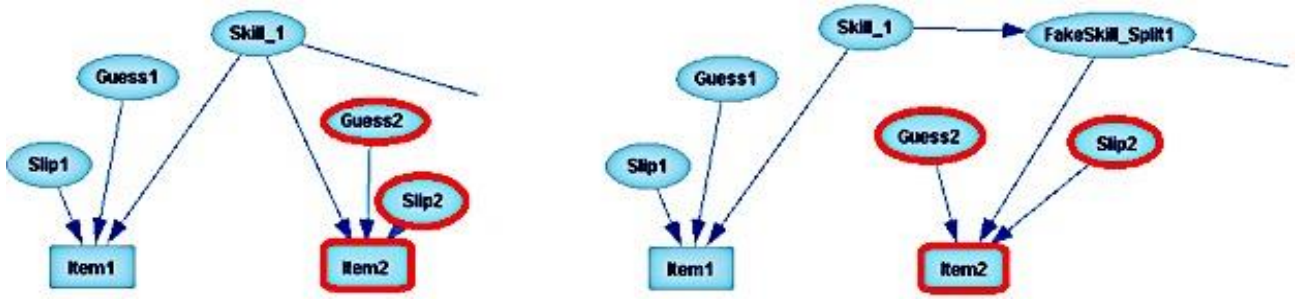


Figure 4-2 Creation of Fake Skill. The left skill graph shows the original skill graph before the creation of the fake skill. The skill graph on the right shows the skill graph after the creation of the fake skill. The fake skill was created from Skill_1 where item 2 was removed from skill 1 and attached to the fake skill

4.3.4 Evaluation

In order to evaluate our Bayesian Network, we used a similar process as done in [1]. We use Expectation Maximization (EM) to learn parameters and fit our model. To evaluate our model, we used per student per item cross validation with 5 student folds and 3 item folds. Our student and item folds were chosen randomly for our evaluation. In [1], the item folds were chosen randomly but kept the same for each student. The only difference between the evaluation in [1] and this experiment is that each student is assigned a different set of random item folds instead of all students having the same set of random item folds.

4.4 Experiments

4.4.1 Experiment 1

In this first experiment, we started with a set of 3-skill graphs. For each of the graphs, we insert a fake skill. We define a fake skill as one that is broken off of an existing skill. The fake skill has a random number of items chosen from the original skill and the fake skill is either a pre-requisite or post-requisite of the original skill. If the fake skill is a pre-requisite of the original skill, all the previous pre-requisites of the original skill become the pre-requisites of the new fake skill and the original skill becomes the post-requisite of the fake skill. The whole idea is to figure out if this fake skill will be easily identified and merged with the skill from which it was created from. This is to validate our merge operations and to determine what factors influence the determination of a better skill-model /skill map than the original.

4.4.1.1 Analysis

We analyzed the results of the experiment and looked at how the number of students, number of items, guess/slip values, and the number of fake skills impacted RMSE of our predictions and the percent of correct graphs learned back. Fig 3-3 shows the relationship between the probability a student guesses/slipped and the RMSE as well as the percent of the correct skill graph being learned back. We paired guess and slip values to lower the number of variables in our experiment. Our guess/slip pairings are as follows: (0, 0), (0.1, 0.08), (0.3, 0.16) and (0.5, 0.25). It shows that the higher chance the student has to guess the answer the less accurate and harder it is to learn back

the true original graph. The percent of graphs learned back with a guess/slip probability of 0 is significantly better than the percent of graphs learned back with a guess probability of .5 ($p < .001$). A realistic guess probability is around 0.14 calculated in [61]. At this point the percentage of graphs learned is somewhere between 0.25 and 0.33. These are not great percentages to learn back a correct graph under realistic guess and slip values. Not much can be done to lower the guess probability on typical questions middle school math students would see. However more student data can be used to increase model performance.

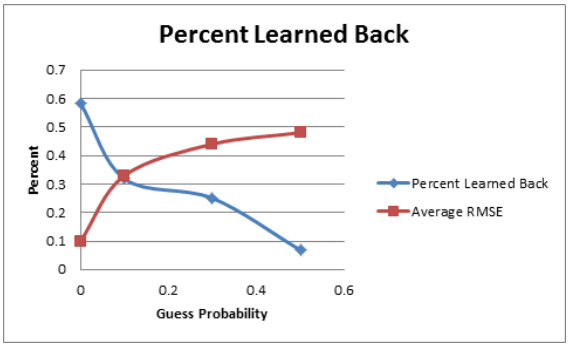


Figure 4-3 Effect of guess/slip on learning back the original graph

The guess/slip probability is the biggest factor that affects model accuracy followed by the number of students. Table 4-2 shows how both the guess/slip probability and the number of students affects the percentage of correct graphs learned back and average RMSE. A cell is broken up into two columns where the first column in the cell is the percentage of correct graphs learned back and the second column in the cell in the average RMSE value.

Table 4-2 Student/Guess Impact on Evaluation

	Number of Students							
Guess	50		100		150		200	
	P_{LB}	RMSE	P_{LB}	RMSE	P_{LB}	RMSE	P_{LB}	RMSE
0	0.33	0.09	0.64	0.08	0.70	0.1	0.67	0.02
0.1	0.25	0.36	0.33	0.33	0.33	0.32	0.38	0.31
0.3	0.25	0.46	0.33	0.44	0.08	0.44	0.38	0.43
0.5	0.08	0.49	0.08	0.48	0.08	0.48	0.00	0.46

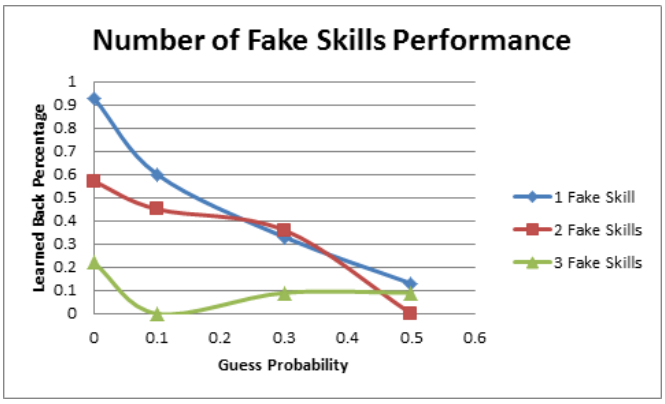


Figure 4-4 Effect of Number of Fake Skills on model improvements

4.4.2 Experiment 2

In experiment 1 multiple randomly chosen graphs were used as the ground truth. In this experiment, we chose to try each possible 3-skill graph to see if the graph structure had an effect on whether or not the correct skill graph was learned back. The methodology was the same as experiment 1 except instead of randomly choosing graphs we ran each of the four graphs for each possible number of students and items per skill. Figure 4-5 shows all four possible 3-skill graphs. After determining that the major factor impacting performance were guess/slip values, a reasonable pair of values were chosen for the guess and slip values (guess=0.1 and slip=0.08). Additionally, we fixed the number of fake skills to one in order to reduce the variability of the factors.

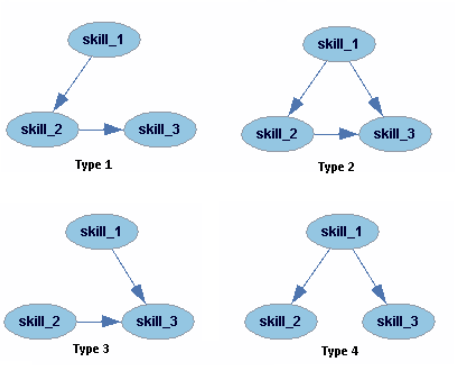


Figure 4-5 Different Graph types for experiment 2

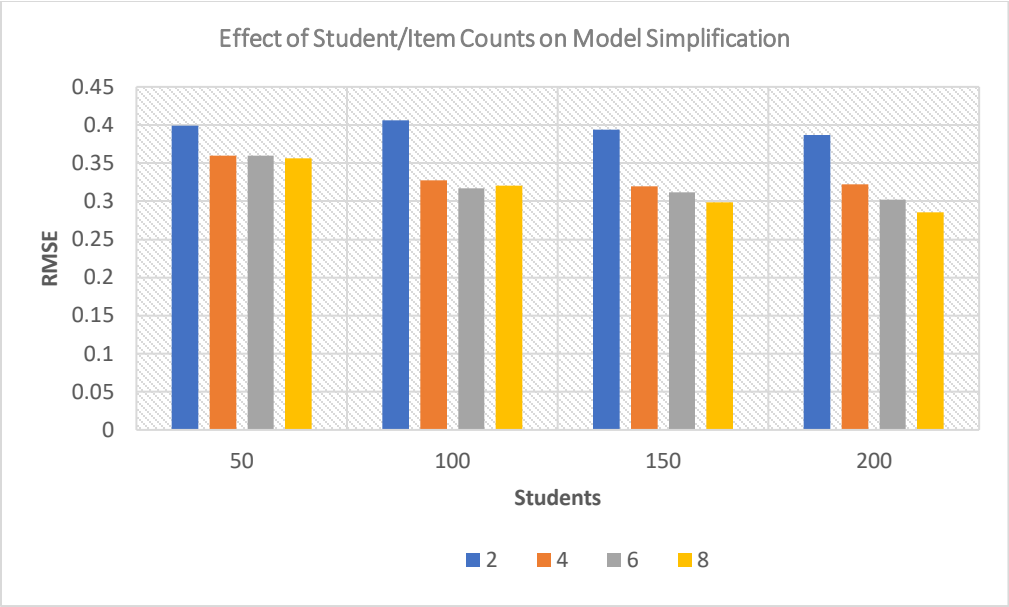


Figure 4-6 Effect of students/items on the model simplification

The general observation from this experiment is clear from Figure 4-6 above. As the number of data points increases, the level of accuracy in recovering the original graph increases. This is in spite of the fact that the location of the fake skill was not fixed. Moreover, for any given number of students, an increase in the number of items results in a slight decrease in RMSE and hence better chance of learning back the original graph. This experiment shows that the data points (i.e. student and item numbers) have an impact on improving on the determination of the best model from a given model.

4.4.3 Experiment 3

In this experiment, we fixed all variables except for the number of students and the number of items per skill. We wanted to see how stable our search was and how well it performed for a small example with reasonable parameter values. We fixed guess at 0.10 and slip at 0.08 with three skills and one fake skill. For the fake skill, we took the first half of items from the original skill. We ran our algorithm for 50, 100, 150, and 200 students for 2 and 8 items per skill. For each pair of parameters, we ran the experiment 10 times with different random seeds and took an average of the number times the correct graph was learned back. Figure 4-7 shows the results of this experiment. We found that the results are very stable for graphs that had two items per skill. The results were less stable for graphs with eight items per skill although the percent of graphs learned back was much batter. The

graphs that had two items per skill were learned back correctly 8% of the time, where graphs with eight items per skill were learned back correctly 43% of the time, which is a significant improvement (n=40, p<.001).

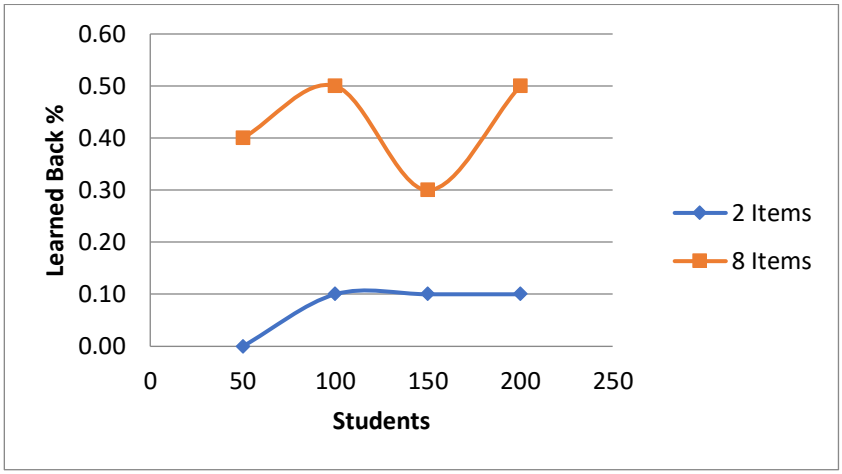


Figure 4-7 Percent of graphs learned back for student ranges 50-200 and 2+8 items per skill.

4.4.4 Experiment 4

We ran experiment 4 to confirm that the number of students has an impact on the recoverability of the original graph, fixing all other parameters at reasonable values and varying the number of students. For this experiment, guess and slip values were set at 0.1 and 0.08 respectively. We used graph type 4 (Figure 4-6), set the number of items to 4 and fake skills at 1, varying the location of the fake skill. The student numbers were varied from 10 to 100. For each student number, the evaluation was run 10 times. The results, in Figure 4-8, show that as we intuitively assumed, the number of students has a huge impact on the algorithm's ability to learn back the true graph. The results show that as the number of students increases the probability of a skill graph being learned back increases whiles at the same time the RMSE reduces. These results, we found, are significant with p-values below 0.01. This finding confirms that student numbers is an important factor that needs to be considered when refining learning maps.

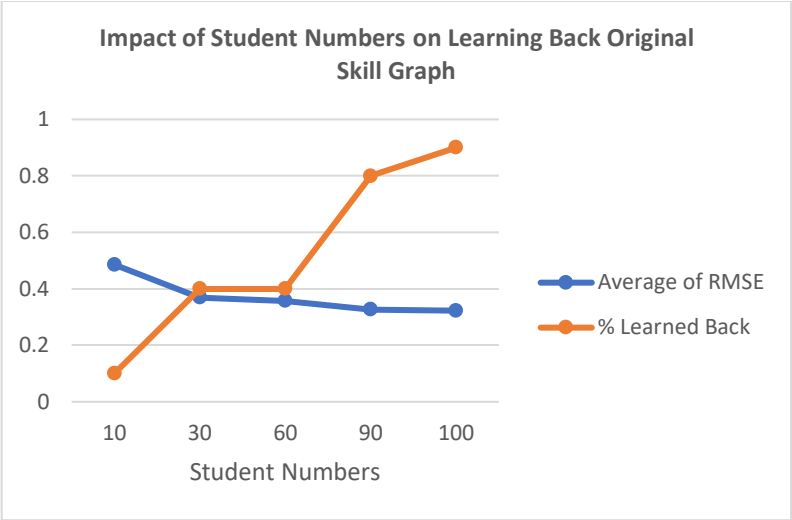


Figure 4-8 Impact of Student Numbers

4.5 Conclusion

Many learning maps/cognitive models are built from expert knowledge. With the production of lots of educational data on student performance, it has become imperative to find data centered methods of improving upon these expert-designed learning maps. In our earlier studies, we designed and presented an algorithm for simplifying/improving the predictive accuracy of these models. In this chapter, we have presented a number of factors that influence the data centered model improvement process we initially published. We have shown with our simulation studies that the guess/slip values, number of items per skill, the number of students and the number of fake skills in the graph affect the simplification of the skill models. We also explored many parameters to see how much data is needed to recover the true learning maps. For future work, we plan to continue to evaluate our algorithm on larger examples to see how well our algorithm can scale up and test it on well-known real data sets.

5 Can Skill Prerequisite Topologies be Accurately Learned Using Deep Knowledge Tracing?

Observing student knowledge over time grants insights into the learning process. Measuring student performance, it often becomes apparent what skills, or knowledge components, compose a student's strengths, as well as potential weaknesses in terms of what content is understood and what requires more practice or remediation. The representation of content into a type of hierarchy is often adopted in the classroom and by computer-based learning platforms as a means of determining the ordering in which skills should be presented to students to maximize learning. This hierarchy, composed of prerequisite links, defines the order in which skills should be learned. While some have been proposed, few quantitative methods exist to verify existing prerequisite graphs. Applying deep learning to this task of predicting next-problem correctness as a student works through an assignment, Deep Knowledge Tracing (DKT) [67] has been suggested to be able to identify latent relationships between skills based on estimates of student knowledge and the ordering in which skill exercises are completed. This work seeks to build upon this previous work by appropriating the original Deep Knowledge Tracing methodology as a means of verifying the prerequisite graph defined with the ASSISTments online learning platform, and adapting that methodology to determine the feasibility of making stronger claims of causality in measuring these skill relationships. It is found through our validation, DKT is an insufficient model to make any claims regarding the relationship of knowledge components without incorporating considerations of skill content.

5.1 Introduction

The ordering of knowledge components, or skills, aligned to a curriculum is often established by domain experts. These orderings, often represented by a hierarchy defining pre-to-post- requisite concepts, however could benefit from data-driven methods of validating that the correct ordering and relationships have been identified. The usage of computer-based systems in classrooms has allowed for the emergence of such methods, but themselves often lack proper validation, motivating our work here. The ability to develop and verify unsupervised methods

of identifying skill topologies, detailing the relationships between skills using evidence from real student data would greatly benefit teachers, administrators, and developers of computer-based learning platforms to implement optimized skill orderings to be presented to students.

Many such models that attempt to identify latent skill structures address the question of whether or not the current skill topologies represent the best sequencing of content for instruction. Can we improve upon the models by identifying missed relationships or removing weak or non-existent links? In fact, the question of learning prerequisite skill structures from learner performance data is beginning to gain traction in the educational data mining and analytics community in recent years. Some methods for learning the skill topology from data have been explored. Among the numerous methods that have been proposed for this exercise is Partially Ordered Knowledge Structures, which was proposed by Desmarais et al. [28]. POKS generates item-level topologies based on the correlation between student performances on items or problems. Adjei and Heffernan [4] used randomized controlled experiments in an adaptive testing system to investigate the strength of relationships between skills in a given knowledge topology, with varied degrees of success. Learning Factors Analysis (LFA) [19] was introduced as another method for refining existing prerequisite skill topologies, and possibly generating new topologies. Pavlik et. al. extended the LFA model in their quest to generate domain models, also referred to as skill topologies. [63] They analyzed learning curves and used the results of their analyses to generate the learning models from student performance data [19]. Additionally, Chaplot, et. al. [20] combined text-based and performance-based methods for inferring prerequisite skill topologies from student performance data as well as from text-based course materials. They proposed an unsupervised approach for the task, and report that this approach outperformed alternative supervised methods.

In recent years, researchers in the learning research community have started applying neural network-based data mining techniques to the task of generating/learning skill topologies from data. Piech et.al. [67] developed Deep Knowledge Tracing (DKT), a neural network approach for predicting student next problem correctness on items in an assignment. The primary goal of that work was to propose a better method of predicting students' performance on the next item in an assignment task, and was meant to be an effective alternative to the widely known probabilistic Bayesian Knowledge Tracing (BKT) model [26]. In spite of this

primary goal, Piech and his colleagues mention that the DKT method produces, as a byproduct, estimates of the relationship between two skills based on student performance on the skills under consideration. They attempted to use these estimates to learn the skill topology represented in the data and hence to improve upon existing methods of learning skill structures from data.

While DKT promises to do well at predicting students next problem correctness, it has its own challenges. Khajah and colleagues [48] find that simpler methods that model student performance, like BKT, perform equally well even though the resources required to run the DKT model far outweigh those required by the simpler BKT models. Also, for the context of learning skill topologies from data, the DKT model generates probability estimates even for unseen skill orderings and does not take into consideration the difficulty of skills when determining the directionality of links. Zhang et al. [100] attempted to replicate Piech's work [67] and made general conclusions about the skill topologies they found, without taking into consideration the limitations just presented. They believed the skill link estimates produced, without checking to see whether those combinations were ever present in the training data. The mere fact that estimates exist for certain skill links does not necessarily mean that those estimates can and should be accepted. Further scrutiny is required to be certain that those estimates presented are reasonable. Additionally, the DKT method for learning skill topologies can easily produce spurious, or non-existent links. As an example, a 9th grade skill may be found as a prerequisite to a 2nd grade skill, which does not make any sense in reality. This therefore requires some human judgements to scrutinize the skill dependency graphs or topologies learned purely from the DKT model.

This work seeks to explore the feasibility of using DKT method to identify causal skill relationships. As described, this has been attempted in other works already, but we believe there are a number of factors that have not been considered in making such identifying claims. To this end, we propose an unsupervised method utilizing DKT to identify prerequisite structures from student performance data.

The following sections present a brief introduction to DKT as a prediction method, as it is used as the basis of our methodology. We then describe the methodology we propose, considering aspects of the skill orderings that should not be ignored. We observe the feasibility of this method using a simulation study and observe the results using real data from the ASSISTments online learning platform as well.

5.2 **Deep Knowledge Tracing**

Knowledge tracing is a task which measures a student's knowledge level on a particular skill given the student's past performance on that skill. Bayesian Knowledge Tracing (BKT) is the most famous model for the knowledge tracing task.[26] BKT models a student's knowledge level on a particular skill as a latent variable, and updates the probability that the student answers the next problem from the same skill correctly using Hidden Markov Models. DKT is recently introduced by [67] and uses Recurrent Neural Networks (RNN) to model student learning process.

RNNs attain the state of hidden nodes and the state summarizes all the information about the past input to the hidden nodes that is necessary to provide an insight to the future prediction. At a given time step, both the input at current time step and the state from the immediate previous time step are fed into hidden nodes. This characteristic allows RNN to memorize the information about the past, which is necessarily useful for the final prediction task. Long short-term memory (LSTM) is a variant of RNN that solves the vanishing gradient problem suffered by RNN. [43, 44] Compared to RNN, each hidden node of LSTM has three gates: input gate, forget gate, and output gate. By controlling these three gates, LSTM have the capability of learning long-term dependencies.

As shown in Figure 5-1, the student performance (correct or incorrect) on a given skill is converted into a fixed length vector. For a dataset with a small number k of unique skills, a one-hot encoded vector with $2k$ dimensions is used to represent the input x_i at time step i. For the first $k$entries in the input vector, one entry is set to 1 if the corresponding skill for that entry is answered correctly. For the other k entries in the input vector, one entry is set to 1 if the corresponding skill is answered incorrectly. The prediction y_i is a vector with k dimensions, where each entry represents the predicted probability that the student answers the corresponding skill correctly.

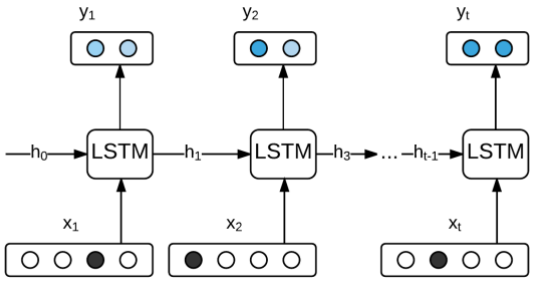


Figure 5-1 An Illustration of DKT. The input ($\boldsymbol{x_i}$) to LSTM is the one-hot encoding, which indicates the skill that is answered at time $\boldsymbol{i}$ and if the skill is answered correctly. The output $\boldsymbol{y_i}$ is a vector representing the probability of answering each skill correctly at the next time step.

5.3 **Methodology**

The methodology used to estimate the strength and directionality of skill links requires careful consideration. Other works have attempted to claim the ability to identify causal relationships inferred from real data [100] without considering other confounding factors. This section describes such factors that must be considered to make stronger claims when inferring causal relationships using predictive models.

The first step of our methodology includes applying the DKT model, as described earlier to our datasets. A simulated dataset and real-student dataset are explored in this work and will be further described in later sections. The DKT model takes a one-hot encoded vector representing correct and incorrect responses for a series of exercises, or skills. As the primary goal of the DKT model is to predict next problem correctness, such predictions are what is output from the model. While individual item predictions can be observed from the output vector, as is done when only observing next problem correctness, the model does output a complete vector of probabilities associated with each skill. To interpret this, once the model is trained, if given an input indicating a correct response on a problem of skill 'A', the output represents the probability that a student answers a problem from each skill correctly. This output answers the question of what is the probability of answering skill 'B' correctly given that the student has just answered a problem from skill 'A'? Considering this probability, in conjunction with the probability of answering other skills correctly given that an incorrect response has just been recorded, has been used in other methods to claim causality. However, two problems emerge from using just this information alone: trusting unseen directionalities and differing skill difficulties.

The first problem stems from the fact that not all skill ordering combinations have been seen within the training data. As students often follow a curriculum, or pre-determined ordering of skills, it is likely that a student may see a skill 'A' before skill 'B' but never in the opposing direction. The DKT model is of course still able to produce an output for this unseen ordering, but it has never been trained to know if that output probability is good or bad; it can only be trained on labels that have been observed in the training data. For this reason, we can only trust those skill orderings that exist in the training dataset. Furthermore, to avoid making claims on just a small set of students, we place a threshold that 60 students must have seen the skill ordering in order to believe the model results. This arbitrary number is chosen to rule out the possibility that just one student, or even one class is the sole instance of a particular skill ordering in the dataset.

The second problem is a much more complex problem, pertaining to differences in skill difficulties. If observing only the probabilities of answering problems in one skill given correctness in another skill, the easier skill is always likely to have a higher probability of correctness. In other words, if a difficult skill is given before an easy skill, it may appear that the probability of answering the second skill is impacted by the first skill even if no true relationship exists. This is partially addressed by using both probabilities of answering the second skill correctly given correctness and given incorrectness on the first skill, but this difference can be impacted by skill difficulty in various different ways. Furthermore, ceiling (and floor) effects become evident when skills are very easy or very hard, where the probability of correctness is always high or low regardless of the existence of skill relationships. As such, we argue that no strong causal claim can be made unless the skills are found to have the same, or at least similar difficulties. Comparing skills on even terms helps to rule out confounding effects incorporated due to skill difficulty.

With this in mind, we use the output of the DKT model where the output skill matches the input skill to measure difficulty. We observe the probability of answering a skill 'A' given that a problem from skill 'A' is answered correctly. For each observable skill ordering, the difficulty as defined by this probability is compared in the form of a ratio of the first skill's difficulty to the second skill's difficulty. If the two values are equal, this ratio is equal to 1, with values farther from 1 indicating large differences in skill difficulties. We further filter the observable skill links to those orderings with a difference ratio between 0.95 and 1.05. While we are filtering, we

are not claiming that no prerequisite links exist in the data we are excluding, but rather that using a method such as DKT for this type of analysis is insufficient to make reliable claims outside our observed space.

With our likely smaller set of skill orderings, we can finally compare the probabilities output from the DKT model. For each observed ordering, we find the difference of probabilities of answering the second skill correctly given that the first skill is correct versus when the first skill is incorrect. For example, if observing the ordering of skill 'A' to skill 'B', we find the difference:

$$P(\text{B is correct} \mid \text{A is correct}) - P(\text{B is correct} \mid \text{A is incorrect})$$

In order to make a claim that A is a prerequisite to B, it would be expected that the difference would produce a comparably large, positive value. This "comparably" large value, however, introduces one last criterion for claiming a causal relationship. As it is not knowable what a "large" difference value is for a particular ordering and it is likely that this value is dependent on the skills within the ordering, the difference value must be compared to the difference value of the opposing skill ordering. For example, the difference value of skill A to skill B is compared to the difference value of skill B to skill A. If both orderings do not exist in the observed set, no strong claim can be made as there is no basis of comparison. This comparison is made for all observable skill orderings with a matching reverse ordered pair, identifying prerequisite ordering to exhibit the larger of the two values. A threshold can be placed to limit the number of believed relationships, which is explored in this work using simulated data.

5.4 **Datasets**

5.4.1 Simulation

In an effort to observe how well the Deep Knowledge Tracing model is able to identify prerequisite relationships, this work includes a simulation study. Using real student data recorded within ASSISTments [42], we begin by modeling individual students. As real student knowledge is a complex problem to model, the real data is used to begin the process. The same dataset used in the other trials of this work is used again here, summarized into a one-student-per-row description of student performance. This performance includes means and standard deviations for percent correctness, the time it takes to complete individual problems, and hint usage within

ASSISTments. As these summary statistics are aggregated over a multitude of different content, each row captures student-level information independent of individual knowledge components. Each simulated student is sampled at random from this dataset, using the same mean and standard deviations of the real student in an attempt to remove some of the artificial nature often exhibited in simulation work.

As we also need to control the difficulty and relationship between skills, 10 arbitrary simulated knowledge components are also created, where the ground truth values are known. Each skill is given a mean and standard deviation representing the difficulty of problems of that content. Difficulty means are sampled from a normal distribution with a mean of 0.7 and standard deviation of 0.2, with higher values indicating a more difficult skill. The standard deviation of skill difficulty for all skills are set at 0.1. In our case, the mean of 0.7 is chosen based on a rounded average of all correctness across all skills in the real data, while the standard deviations are chosen arbitrarily to observe a controllable range of difficulties.

With these simulated skills, a set of relationships are also defined. The first set of relationships define the effects of practice within a skill. Essentially, particularly in mastery-based assignments, it is assumed for simulation that students who demonstrate understanding of concepts are more likely to answer problems of the same content correctly in the future. This effect is stored within memory of each student, as it is likely in real scenarios that different students will sometimes see differing amounts of content; when a student answers the first three questions correctly, for example, the assignment is finished and that student is likely to have seen less of a range of content than a student needing 8 problems to complete. This effect of practice is sampled from a normal distribution with a mean of 1 and standard deviation of 0.02, and represents a scaling to student knowledge when observing similar content in future simulated assignments. This effect incorporates a tendency to slip, as the scaling factor can sometimes drop below 1; this aspect is included as it is not always the case that students improve with more practice. It is also the case that we do not model time differences between assignments, which may also impact student memory when seeing similar content after an extended period of time. While this effect impacts student knowledge, the relationship also impacts the amount of time students take on future problems of the same content and hint usage within that skill, with these being set arbitrarily at 0.98 (sd=0.1) and 0.9 (sd=0.02) scaling respectively. It is important to mention that for this work, unlike knowledge and hint usage, the

time needed to solve problems does not impact correctness but is included for robustness as this method of simulating data is among the contributions of this work and is usable for future works.

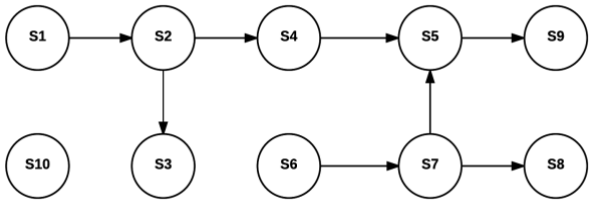


Figure 5-2 Simulated Graph

In addition to each skill being related to itself, a prerequisite hierarchy is also defined, as seen in Figure 5-2. Note that the arrows in this graph and subsequent graphs point from the prerequisite skill to the post-requisite skill. As an example, S1 → S2 implies that S1 is the prerequisite skill of S2. Also, note again that arbitrarily constructed, the graph illustrates several possible structures, including skills having multiple prerequisites (S5), multiple post-requisites (S2 and S7), no post-requisite skills (S3, S8, and S9), and a skill that has no relation to any other skill (S10). Similar to the relationships defined for the effects of practice, the prerequisite effect on knowledge is sampled from a normal distribution with a mean of 1.3 and standard deviation of 0.1. As a prerequisite relationship is clearly defined to be that a skill 'A' impacts performance on a skill 'B', the knowledge effect scaling is given a lower bound of 1.1, ensuring that the effect is never negatively impactful on performance in the post-requisite. Scaling factors for speed and the probability of using hints is also set as constant for all skills at 0.95 (sd=0.1) and 0.95 (sd=0.02) respectively. The assumption is made that hint usage may not be as impacted in prerequisite relationships as they are in practicing the same content due to the differences in material

Assignments are constructed from the simulated skills and given to each of the simulated students. For this particular work, 1000 students are created and each is given 6 assignments. Each assignment is composed of multiple skills, with one skill appearing with higher probability than the others. In practice, it is common for teachers to assign work that is composed of one or two content types, depending on what material has been covered in the classroom. This is apparent as even in the real data, students have opportunities to see certain skills in different orderings. This is of course not randomized, and usually follows a curriculum, so the same is

done in our simulation. Using the simulated graph, assignments follow the pre-to-post ordering of skills, with only small probabilities that a student will see a post-requisite item in the same assignment as its prerequisite.

The simulated students are given problems of particular skills on the assignment, with difficulties sampled from the distribution defined earlier. After applying scaling effects based on previous completed assignments and the skill tagging of each problem, an answer is generated from the student's knowledge distribution. The problem is considered correct if the student answer is larger than the problem difficulty. As hint usage in ASSISTments marks the problem incorrect, the simulated correctness value of each problem is modified in the same way. The student's propensity to ask for a hint, as sampled from the real data, is used to determine the probability that a student asked for a hint on the given question, marking the problem as incorrect if a hint is used. To better model student knowledge over time, the propensity to ask for hints is reduced in scenarios of consecutive correct responses; in other words, the simulated students require less help as they demonstrate understanding of the material. Students are given problems until either answering three consecutive problems correctly, or a maximum threshold of 10 problems is reached. This arbitrary threshold is provided as student persistence is not included in the model.

The resulting simulated performance is formatted in the same manner as the real data, detailing each student's sequence of responses accompanied by the skill tagging of each skill. The resulting summary statistics, as compared to the real data, is detailed in Table 5-1. The resulting simulated data is shown to exhibit similar metrics to that of the real data. This similarity supports the claim that the simulated student performance sufficiently models the real-world data. With all ground truth values and relationships known, the methodology can be compared for its effectiveness. As prerequisite relationships in real data is often difficult to measure, the complexities introduced for simulation will help gain an understanding of how effective this methodology is able to learn latent structures in the presence of simulated noise.

Table 5-1 Summary statistics for generating simulated data

Feature	Real Student Data	Simulated Data
Percent Correct	0.72	0.76
Average Problems per Assignment	5.09	5.4
Hint Usage	0.15	0.14
Time Per Problem (seconds)	24.12	26.15
Completion	0.8	0.82

5.4.2 ASSISTments Dataset

ASSISTments is an online homework assignment and completing system that has served a number of students and teachers for about a decade now [42]. A recent study showed that this system causes a huge gain in learning among students of different knowledge levels as compared to students who did not use the system. The study thus shows the effectiveness of the system in helping students learn over a given period of time. [79] The current users of the system span all over the US with a heavy concentration of the schools in the north-eastern parts of the USA. A vast majority of the student users are at the middle school level, with a few of them being in the high school. The system provides teachers with the ability to assign sets of questions, referred to as problem sets, to students. The problem sets are usually composed of problems that are tagged with skills/knowledge components and are presented to the students in random order. Students continue to answer the questions in the problem set until they get a predefined number of questions correct in a row. This predetermined number of questions is referred to as the mastery criterion. Daily limits are set for students who are unable to complete the assignment task on that day. These limits define the maximum number of questions by which the system should stop presenting problems to the students if they have not reached the mastery criterion.

We pulled data from this system for the 2014-2015 academic year. The complete data set for the academic year consisted of about 5 million unique rows, each row representing a student's performance on a problem in a given assignment. From this large sample space, the first one million rows are selected to perform our analysis,

utilizing a large sample space in which to apply and test our method while at the same time considering computational costs. This subset of the data set consisted of student performance data for 25,465 students whose grades range from 4 to 9. There was a total of 124 knowledge components in the dataset. The students had completed an average of 40 problems for all assignments in the dataset. A total of 32,465 unique assignments had been attempted. The complete dataset as well as the code used for the current study are available at http://tiny.cc/DKTPrereqSkillsGraph.

5.5 **Results**

The results of our method are presented here, illustrating the learned skill topologies. We are able to evaluate our methodology, using the results of the DKT model and considering confounding effects caused by skill orderings and differences in skill difficulties. To reiterate the purpose of this evaluation, the use of a predictive model to make causal inferences seems inherently problematic. Therefore, using such a method and accounting for impactful effects should give us an idea of the feasibility for use on this task, particularly when evaluated on the simulated dataset where ground truth values are known.

5.5.1 Simulation

As depicted in Figure 5-2, we defined a ground-truth prerequisite structure for use on our simulated data. The simulation, proving to illustrate in Table 5-1, sufficient representation of real world data on which it is based, uses the defined structure to emulate students solving problems. From that data, the skill graph seen in Figure 5-3 illustrates what our method determines to be the skill topology.

In this learned graph, the method is able to identify just three direct prerequisites in skills 1 to 2, skills 4 to 5, and skills 5 to 9, while identifying other skills that exist within the same prerequisite branch, such as skills 1 to 3 and skills 4 to 9. While illustrating some success, the method also identifies several spurious links, particularly those involving skill 10, which has no true relationship to any other skill. If existing in a simulated study that, while close to real student sequences, is likely to capture only a portion of the true complexity, such limitations are likely to exist if applied to real data; this occurs even in the case of accounting for differences in skill difficulty and attempts to limit our comparisons to those where causal claims are strongest.

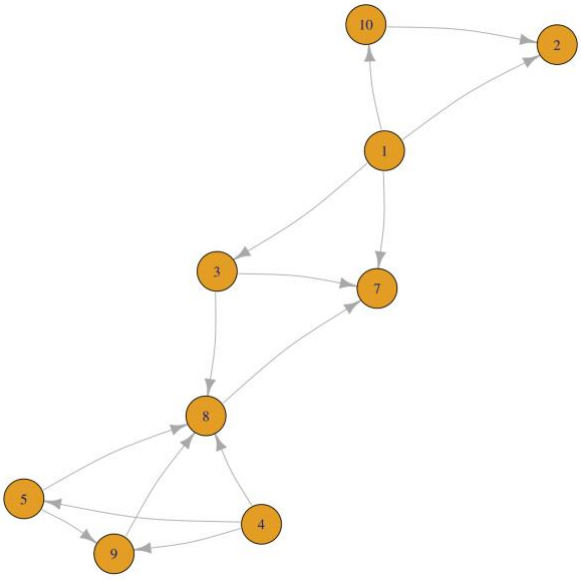


Figure 5-3 Learned Graph from Simulated Data

5.5.2 Real Data

The findings of our simulation study suggest that, while able to correctly identify some prerequisite relationships, DKT and out method is prone to identifying false positives, or spurious links that are truly non-existent. The method is not completely invalid, however, suggesting that while perhaps failing as an unsupervised method, incorporating expert knowledge to supervise the believability of identified links may improve its accuracy. The model may be useful, therefore, in a human-assisted manner to further base identified relationships within the interpretability of the knowledge components themselves.

In this regard, we apply our method to the real-world data and observe the results in Figure 5-4. Notes that Table 5-2 contains the list of skills that correspond to the identifiers in the nodes of the graph. Those reported underwent two levels of human-guided filtering before being considered. The first is an additional constraint placed on the links to help remove likely spurious relationships. Using the common core standard tagging of each of the skills, only those relationships in which both skills are in the same grade are considered. This helps remove the possibility that a 9th grade skill could be identified as a prerequisite to a 2nd grade skill. The remaining links are illustrated in the figure. From these links, a human domain expert can scrutinize the identified relationships further.

Figure 5-4 shows the resulting prerequisite skill graph that was inferred from the results of the DKT method together with our intuitive method for determining the direction and strength of the prerequisite skill links.

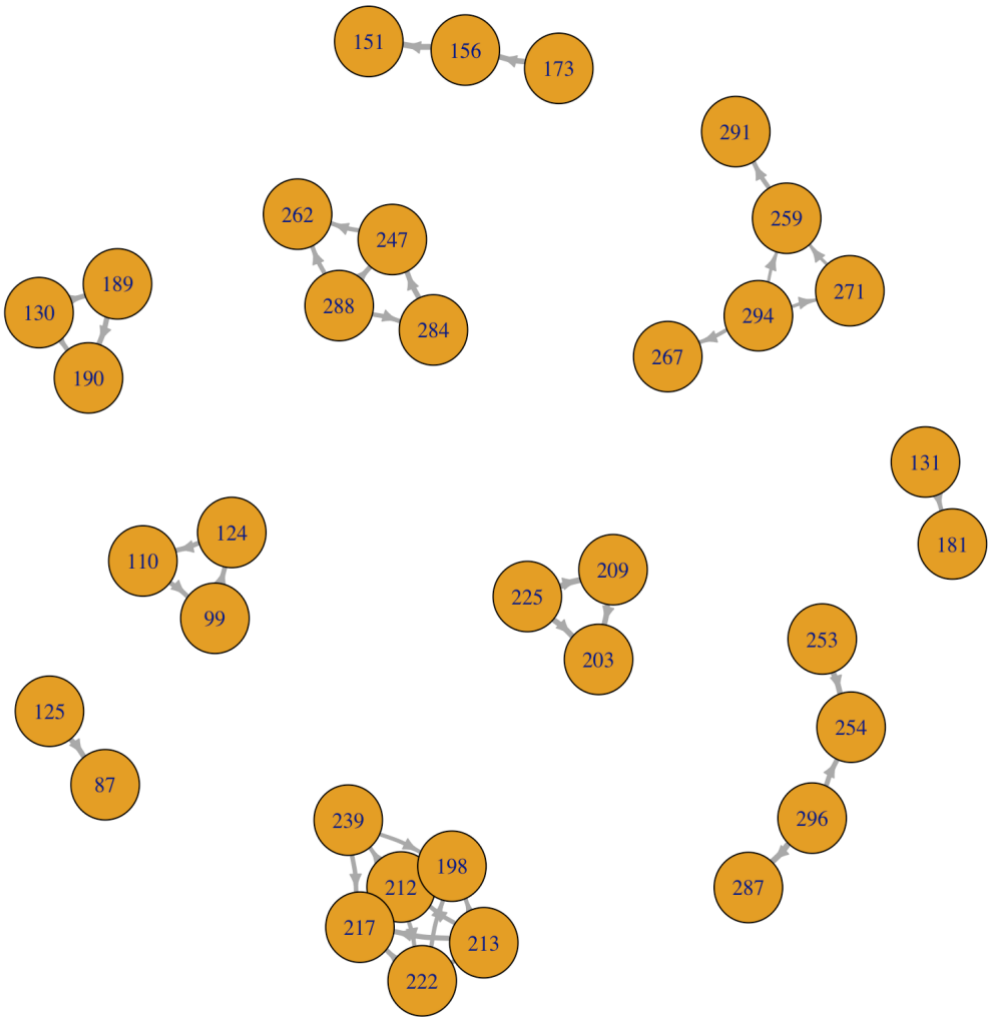


Figure 5-4 Learned Graph from Real Data

Table 5-2 List of skills with their corresponding ids from ASSISTments

Skill ID	Common Core Code	Skill Name
87	5.NBT.A.3b	Comparing Positive Decimals
99	5.NF.B.4a-1	Multiplying Fractions
110	5.MD.A.1-2	Unit Conversion Within a System - Standard
124	5.G.B.4-1	Properties and Classification of Polygons with 5 or more sides
125	5.G.B.4-2	Properties and Classification of Quadrilaterals
130	6.RP.A.3b	Unit Rate
131	6.RP.A.3c-1	Percent of
151	6.NS.C.7a	Comparing Integers on a Number Line
156	6.EE.A.1	Evaluating Exponents
173	6.G.A.1-1	Area of Triangles
181	6.SP.B.4-2	Box and Whisker
189	6.SP.B.5c-2	Mean
190	6.SP.B.5c-3	Median
198	7.RP.A.3	Percent Word Problems
203	7.NS.A.1d-1	Adding Integers
209	7.NS.A.2c-2	Dividing Integers
212	7.NS.A.3	Integer Word Problems - All Operations
213	7.EE.A.1	Combining Like Terms
217	7.EE.B.4a	Equation Solving - Two or Fewer Steps
222	7.G.B.4-3	Circumference
225	7.G.B.5-1	Complementary and Supplementary Angles
239	7.SP.C.7a	Probability of a Single Event
247	8.NS.A.2-1	Approximating the Square Root
253	8.EE.A.3	Scientific notation
254	8.EE.A.4	Operations with Scientific Notation
259	8.EE.C.7b	Solving Linear Equations
262	8.EE.C.8b	Systems of Linear Equations
267	8.F.B.4-1	Finding Slope in an Equation
271	8.F.B.4-5	Writing Linear Equations from Ordered Pairs
284	8.G.A.3-1	Translations
287	8.G.A.4	Identifying Similar Triangles
288	8.G.A.5-1	Angles Formed by Parallel Lines and Transversals
291	8.G.B.7	Pythagorean Theorem
294	8.G.C.9-2	Volume of a Sphere

Observing the skill orderings identified by the method in Figure 5-4, it must be noted that the groups of skills do not necessarily represent clusters of skills, especially since these groups of skills do not represent sub-domains. They just represent skills that our method identified as related directly or indirectly to each other. Also, several interpretability issues become prevalent. First, only one link (189 → 190) is identified to match with the existing skill structure of 157 links between the 124 observed skills within ASSISTments. While it is among the goals of such a model to identify links that may have been missed by domain experts who constructed the prerequisite hierarchy, identifying only one to match domain experts raises concern that the model is not accurate. It is therefore more likely that the model is identifying many more spurious than actual prerequisite relationships in the real data. Furthermore, the method produces a cycle between the skills 99, 110, and 124; while a method could be introduced to break cycles, the emergence of such an occurrence further suggests that the relationships

identified are not prerequisite in nature. The latent structures do not pertain to content, but rather are largely impacted by other emerging correlations in the data.

5.6 **Discussion**

The disappointing validation results of our simulation study suggest that a predictive model such as DKT is insufficient to make reliable causal claims regarding skill orderings. Even when accounting for confounding effects caused by unseen skill orderings in the data and differences of skill difficulty, the model identifies many spurious links. This is not to say that the model is not at least partially effective; it was able to identify three prerequisite relationships and other links within the same branch of the hierarchy. However, as these links were indistinguishable from the spurious links identified, there is no known unsupervised way to know which links are believable when applied on the real data.

Aside from the ability to make causal claims, it is similarly unreliable to make correlational claims for the same reason of spurious links. Based on the simulation study, skill 10, which is defined to have no relationship whatsoever to other skills, is identified as having relationships with two skills. It is believed that this is the case due to confounding effects of the skill difficulties once more; unlike the problem of differences in skill difficulties, this problem actually stems from similar skill difficulties. If the model is identifying a relationship, it is possible that the skills are only correlated in terms of the difficulty of problems within that skill, but not on the content itself, especially since this study did not consider the content of the knowledge components. The DKT model, as presented in Piech et al. [67] has not been developed to capture aspects of content, and instead observes an imperfect representation of content by means of student performance. If skills have the same difficulty, especially in the case of very high or very low likelihoods of answering problems correctly, the ceiling effect introduced confounds the results leading to such spurious links. For example, in an extreme case, if students answer 100% of problems correct in skill ‘A’ and the same percentage correct in skill ‘B’, the two skills are likely to be identified as related even if there is no similarity of actual content.

5.7 **Contribution**

As is found from our attempts to remove effects of skill difficulty and the relationships identified by the simulation study, it is suggestive that the difficulty of knowledge components is poorly representative of their content. We argue as a result that models observing only this difficulty, such as in the form of propensity of answering items correctly based on other correctness probabilities, is insufficient to make claims, causal or otherwise, regarding the relationship between skills. If employing such models, information of the domain and content should be observed in order to properly support any such claims.

Understanding the problem however, simply motivates future work to develop and utilize models that are able to better represent and utilize content rather than relying on student performance alone. Incorporating such models would exhibit more promise in its ability to distinguish true prerequisite links and non-existent, spurious links.

It is also likely more beneficial to conduct randomized controlled experiments to observe the effects of skill orderings. Doing so, however proves to have its own limitations due to ethical implications of potentially presenting students with a detrimental skill ordering, but some systems have developed means of doing so using remedial work rather than for initial instruction. Coupling such trials with unsupervised methods considering the content of knowledge components is topic worthy of future research.

6 Predicting Student Performance on Post-requisite Skills Using Prerequisite Skill Data

Prerequisite skill structures have been closely studied in past years leading to many data-intensive methods aimed at refining such structures. While many of these proposed methods have yielded success, defining and refining hierarchies of skill relationships are often difficult tasks. The relationship between skills in a graph could either be causal, therefore, a prerequisite relationship (skill A must be learned before skill B). The relationship may be non-causal, in which case the ordering of skills does not matter and may indicate that both skills are prerequisites of another skill. In this study, we propose a simple, effective method of determining the strength of pre-to-post-requisite skill relationships. We then compare our results with a teacher-level survey about the strength of the relationships of the observed skills and find that the survey results largely confirm our findings in the data-driven approach.

This chapter is published at the following venue:

Adjei, S. A. & Heffernan N.(2016) Predicting Student Performance on Post-requisite Skills Using Prerequisite Skill Data: An alternative method for refining Prerequisite Skill Structures. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, Edinburgh, United Kingdom — April 25 - 29, 2016 pp469-473

6.1 Introduction

Prerequisite skill structures represent the ordering of skills in a given knowledge domain. The learning sequences represented in prerequisite skill structures have become an area of interest over the past few years. As a prelude to the objective of learning prerequisite skill structures from data, Tatsuoka developed and proposed the Q-Matrix, a structure that represents the mapping of items on a test to specific skills. [86] Others have built on this structure to find relationships between the skills and items represented in the Q-matrices [12, 82], or proposed methods for refining Q-Matrices [32]. Brunskel presented preliminary work in which she used students' noisy data to infer prerequisite structures. [16] Additionally, Scheines, et al. present an extension of a causal structure discovery algorithm in which the assumption of pure items is relaxed to reflect real data, and use that relaxed assumption to infer prerequisite skill graphs from students' response data.[82]

The focus of other researchers in the community has been on refining the prerequisite structures developed either by domain experts or through data mining approaches, as used by Barnes. Barnes, 2005 #28} Cen, et al. proposed Learning Factors Analysis (LFA) as a method for refining cognitive models. Their approach includes statistical techniques, human expertise, and combinatorial search to refine cognitive models. Following the proposals made by Cen et al. in [19], Adjei et al. [1] developed a combinatorial search algorithm based on LFA and found simplified prerequisite structures, which have equally good predictive power as the originals.

Desmarais, et al. introduced a method for determining partially ordered knowledge structures (POKS) from student data. [31] The main idea behind this approach is to compare pairs of items in a test in order to determine any interactions existing between each pair. The interactions serve as a basis for determining the relationship between the skills represented by the items. Pavlik and his colleagues applied POKS to analyze item-type covariances and proposed a hierarchical agglomerative clustering method to refine the tagging of items to skills, [64] and later proposed Learning Factors Transfer Analysis [63] as a means for generating domain models. Adjei and Heffernan used randomized control experiments to identify links within prerequisite skill structures that require further scrutiny. [5] All of this effort that has been expended in the quest to find skill structures from data have yielded varied degrees of success.

The desire to find the best representation of skills (i.e., the prerequisite skill structure) is important for a number of reasons. It informs domain experts about the optimal sequencing of instruction in order to achieve the best tutoring for students. Additionally, this should help researchers in the education research community to better model students' knowledge and performance in intelligent tutoring systems more accurately. Such strategies and models can benefit students' understanding of new skills by supplying them with the optimal foundations for the material. Likewise, better student models can lead to improved intervention design for those students requiring further aid.

This current study proposes a simple method for identifying problematic links in a prerequisite skill structure, pointing domain experts to the ordering of instructions that may be creating problems for students. In this study, we use linear regression of students' performance on items presented to students in the order of a given prerequisite skill structure and make suggestions about the strength of the relationships between the skills.

This chapter starts out by describing PLACEments, the adaptive testing system from which data was collected for use in this study. This is followed by a description of the methods we employed and the results of the studies. We then present the results of a teacher survey that we conducted and compare the results of the survey with the findings of our data mining task. The chapter concludes with a discussion of the results and possible future work in this area.

6.2 **PLACEments**

PLACEments, a free mathematics adaptive testing system, is a feature of ASSISTments (a free web-based Intelligent Tutoring System (ITS)). When assigning a PLACEments test, an initial set of skills are selected for the test. Students are tested on the initial set of skills and depending on their performance, the system traverses a skill graph to present problems from the prerequisite skills of the initial set of skills. The test adapts to the student's performance as well as the underlying prerequisite skill graph. If a student performs poorly on an item in the test, they are presented with items from the prerequisite skills required to solve the original problem. PLACEments uses a prerequisite skill structure created by one of the experts who developed the Common Core Standard for mathematics. [17] Portions of this structure are currently being used by websites like AchieveTheCor.org [http://www.achievethecore.org/coherence-map/]. The developers of the site call it the Coherence Map.

PLACEments has an additional feature that assigns remediation assignments to students who perform poorly on a test. These remediation assignments are intended to build the students' understanding of the skills they performed poorly on, during the test. The remediation assignments are released in the order of the arrangement of skills in the prerequisite skill structure. Students are assigned lower grade level prerequisite skills first, and until they complete those remediation assignments, post-requisite skills-related remediation assignments are not released. This ensures that the students gradually build on their knowledge of skills until they eventually reach a desired level of mastery of the skills in the given domain.

To illustrate how PLACEments works, Figure 6-1 shows a hypothetical prerequisite skill graph where the letters A through H each represents a skill. The graph additionally shows a typical configuration of a student's

navigation through the prerequisite skill structure in the process of taking a PLACEments test. "Dividing Positive Decimals", "Greatest Common Factor" and "Least Common Factor) are the initial skills assigned on the test.

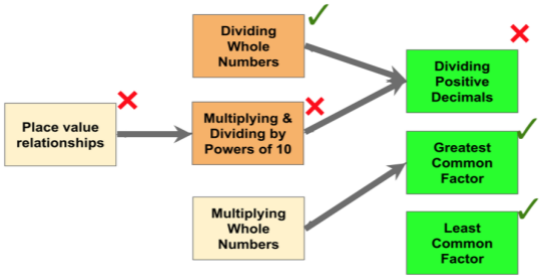


Figure 6-1 A Typical Student's navigation in PLACEments

In this case, the student answered incorrectly the question related to "Dividing Positive Decimals" and so is asked questions for "Dividing Whole Numbers" and "Multiplying & Dividing Powers of 10". Since the student could not demonstrate understanding of "Dividing Whole Numbers", he is further asked questions from "Place Value Relationships", which he performs poorly on as well. PLACEments creates remediation assignments for each of the skills the student performs poorly on (Dividing Positive Decimals, Multiplying & Dividing Powers of 10, and Place Value Relationships). For this particular example, the remediation assignment for "Place Value Relationships" is released before any other remediation assignments are released. The assignment for "Multiplying & Dividing Powers of 10" is released after the student completes that previous skill's assignment.

For the purpose of this study, we focus only on the remediation assignment management feature. This is the feature that provides us with data for determining how strong prerequisite skill relationships are. The remediation assignments are typically assignments in which students practice a number of similarly designed problems to help them master a particular common core skill. In the course of the assignments, students are allowed to ask for help (in the form of hints) as they progressively answer the questions. The student is deemed to have mastered the skill if he/she correctly answers *n* consecutive problems in the assignment without asking for hints. The value of *n* typically ranges from three to five depending on the designer of the problem set. If after a set number of problems (typically called the daily limit), the student is unable to reach the mastery criterion, the system pauses the practice session until the next day when the student can continue with the assignment.

6.3 **Methodology**

6.3.1 Dataset

The remediation assignment feature of PLACEments served as the source of data for the current study. The dataset includes students' performance on remediation assignments. There were 495 prerequisite skill links from the prerequisite skill structure described above. In this study, we focused our attention only on skills that have exactly one prerequisite skill, but it is important to note that our approach is not inherently limited to such skills. Of the 104 skills that have exactly one prerequisite skill, we had 24 of the links that had data for a minimum of 50 students. For each of the prerequisite skill links examined, there was an average of 120 students who were assigned remediation assignments of both the prerequisite and post-requisite skills of the link.

Each row in the dataset has a student's performance on the pre- and post-requisite skills (measured by the percent correct of the Skill Builder, and the number of items it took them to complete the Skill Builder typically referred to as the student's mastery speeds) and the student's prior performance on all problems in ASSISTments. The latter is to help us account for the student's knowledge level. The data set also includes the skill difficulty values for both the pre- and post-requisite skill. These difficulty values are the percent correct for all the items tagged with that skill in ASSISTments. Table 6- 1 shows a sample of the dataset that was used for this study. Each row in the dataset represents a student's' performance on the remediation assignments related to a given PLACEments test. If the student had a similar pair of assignments in another PLACEments test, that information was ignored because we did not want to duplicate the data for a given student. In all, the dataset had 5803 instances of student's performance on pre- and post-requisite skills, involving 1567 students who have completed PLACEments tests.

Table 6-1 Sample data set. [1]

SID	PsSk	PreSk	PosMS	PreMS	StPr	Pre Dif	Pos Dif
23412	57	50	4	5	0.75	0.32	0.40
24321	87	50	3	5	0.86	0.58	0.67
….	…	…	…	…	…	……	.....

6.3.2 Regression Models

We ran linear regression to predict students' performance on the post-requisite skill. To avoid bias caused by student performance and differences in skill difficulty, we included each student's prior performance, mastery speed of the prerequisite skill, and the difficulty of both the pre- and post-requisite skills into our models. The dependent variable was the mastery speed of post-requisite skills.

The following equation illustrates the regression model learned from the data for each of the links:

$$m_{i,j} = \alpha_i + \beta_k m_{k,j} + \gamma_{i,k} K_j + \rho_k d_k + \sigma_i d_i \ldots\ldots\ldots\ldots (1)$$

where i indicates the metric for the post-requisite skill and k indicates the prerequisite for student j. The term m represents mastery speed, α represents the intercept, K represents the prior knowledge, and d represents skill difficulty. β, γ, ρ and σ represent the coefficients of the independent features in the regression.

We considered a link's model only when the model was found to be statistically significant ($p<0.05$) with R-Square above 0.1. All those models with R-Square values below 0.1 were considered to be suggestive of non-existence of a believable link between the two skills. For the models that met the above criterion, a prerequisite relationship was considered to exist when there is a positive standardized beta coefficient for the prerequisite skills mastery speed (i.e., $\beta_k > 0$) and is significant ($p<0.01$ in many cases and $p<0.05$ in a few).

[1] The complete dataset can be found at http://tiny.cc/mslinkstrength. SID is the unique student identifier, PsSk is the post requisite skill id, PreSk has the prerequisite skill id, PrMS and PosMS contain the student's mastery speed of the pre- and post-requisite skill respectively, StPr is the students' prior percent correct (an indication of the student' knowledge level), and PreDif and PosDiff is the difficulty of the pre- and post-requisite skills. The column names have been shortened for lack of space.

Since outliers in the dataset could skew the results, we used two data transformation methods to minimize the effects of outliers in the dataset. The first method was to winsorize the mastery speeds in which all mastery speeds above 10 had their values set to 10. Skill Builders in ASSISTments have this feature where a daily limit of 10 is set to prevent students from banging their heads when they are unable to master the skill within 10 opportunities. This is the reason we chose 10 as the cut off number in order to fairly account for student performance. More than 80% of the data we used had mastery speeds below 10 so the impact of this transformation was not very significant. The second data transformation method we used was a log transform of the mastery speeds. We then used each of the transformations to predict the correspondingly transformed mastery speeds and present both results in the results section.

In the case of the transformed data, we replaced the raw mastery speeds in the model with the transformed mastery speeds. We run linear regression models similar to equation (1) above with the mastery speeds, $m_{i,j}$ and $m_{k,j}$, respectively replaced with the transformed data, $\overline{m_{i,j}}$ and $\overline{m_{k,j}}$. Equation (1) in this case becomes:

$$\overline{m_{i,j}} = \alpha_i + \beta_k \overline{m_{k,j}} + \gamma_{i,k} K_j + \rho_k d_k + \sigma_i d_i \quad \text{..... (2)}$$

6.3.3 Teacher Survey

To verify the results of our findings, we ran a survey of 45 randomly selected domain experts and teachers who use ASSISTments and asked about their perceptions of the strength of the 24 prerequisite skill relationships, including the 14 links we studied in the regression study. A sample survey question is shown in Figure 6-2. The survey had 26 different questions, the first question introduced the survey and the last was complimentary. There was a survey question for each of the 24 prerequisite skill links. For each prerequisite skill link, we presented a sample problem for each of the post- and pre-requisite skills and asked teachers to rate, on a scale of 1 to 7 (1 not important; 7 extremely important), how important it is for a student to know the prerequisite skill to be able to answer the problem from the post-requisite skill. Even though the questions give the impression that we are trying to figure out how related the skills are, we intentionally did not use the terms "pre-" and "post-requisite skills" in order not to confuse the respondents, or to point them in a particular direction. A link is considered to exist if the

mean of the responses for that link was approximately 5 and a standard deviation of less 1 or less. We then compare the results of the survey with the findings of the study and report on the comparison.

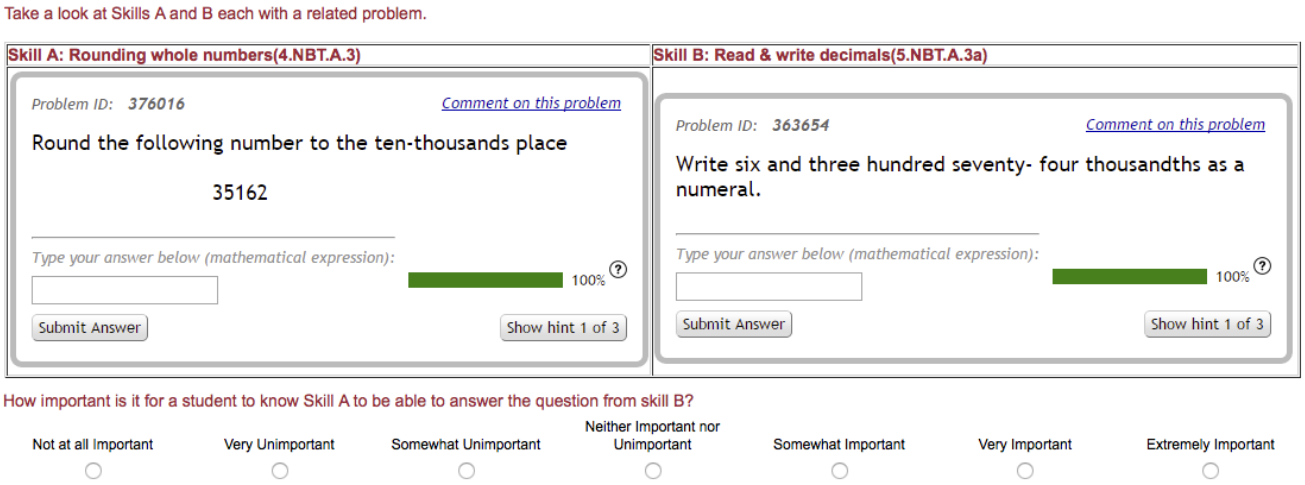


Figure 6-2 Sample Teacher Survey question

6.4 **Results**

6.4.1 Regression

The results of the regression study are illustrated in Figure 6-3, which shows that several of the links could be found to be problematic and require further scrutiny. When the mastery speeds were transformed to take care of the outliers, the models do a better determination of the good links than was the case when the raw speeds were used. The takeaway from this graph is that we do a better job at finding both good and bad links in a prerequisite structure (and thus refine the structure) when we transform the mastery speeds. By transforming the data, we increase the good links by about 10 percentage points, as shown in Figure 6-3.

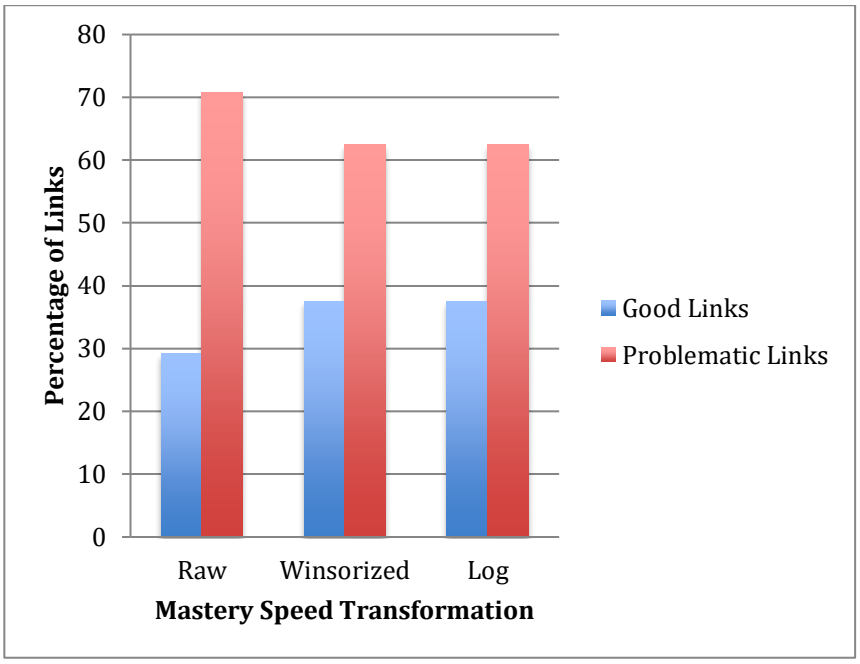


Figure 6-3 Percentages of identified good and problematic links based on mastery speed transformation methods

The bar chart in Figure 6-4 shows that, of the twenty-four (24) prerequisite links, the regression method identified 25% of the links (6 links) in which the students' performance on the prerequisite skills significantly predicts their performance on the post-requisite skills, irrespective of the data transformation method used. Thirty-six percent (36%) of the prerequisite skills (in 8 links) are significant predictors (p-value < 0.05) of students' performance in the post-requisite skill related assignment, if we used any two of the transformation methods. This suggests that a prerequisite skill relationship truly exists between the two skills in each of those links.

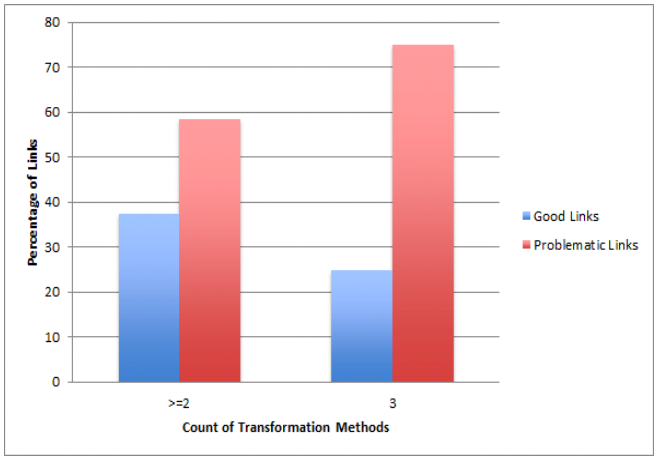


Figure 6-4 Agreement between mastery speed transformation methods

6.4.2 Teacher Survey

We received responses from 21 of the 45 teachers invited to respond to the survey, representing a response rate of 47%. All respondents completed the survey in its entirety, and the resulting scores were averaged per link. Those links found to have an average score greater than or equal to 5 with a standard deviation approximately equal to or less than 1 were viewed as exhibiting a prerequisite relationship. This is concluded as we used a 7-point scale, with those scores greater than 4 indicating at least some importance for one skill to be presented before the other. On the basis of these criteria, the survey found 67% of the links (16 links) are good, while the remaining 33% (8 links) are bad.

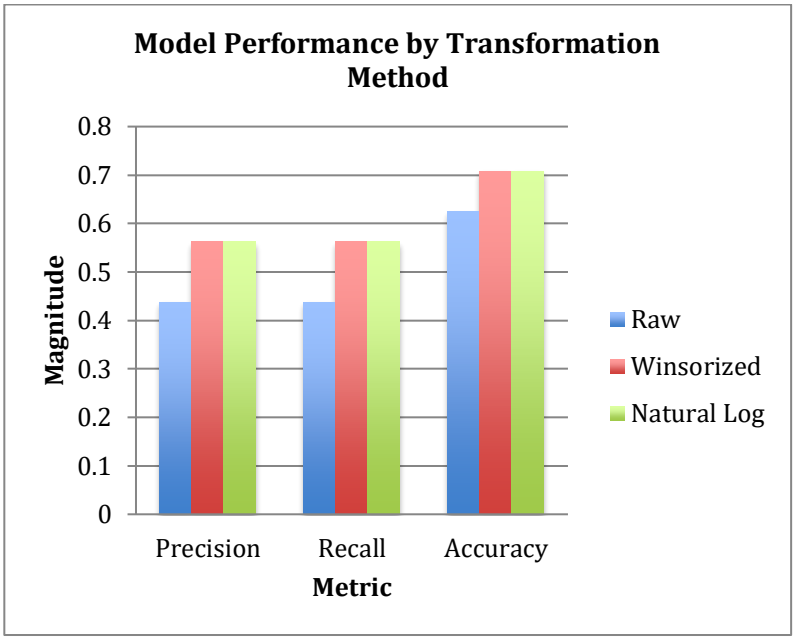


Figure 6-5 A comparison of model predictions with teacher survey about link strength.

Using these values as ground truth, we compare the results to our regression models. In Figure 6-5, we observe our models in terms of precision, recall, and accuracy metrics against the ground truth values. An interesting finding from those results is that the accuracy of the regression models was not influenced by the transformation method used. This could indicate that the transformations alter the data in similar ways, or simply that there were too few instances affected by the transformations to observe an effect.

Figure 6-6 illustrates each method's ability to identify correct links when compared to the ground truth values. We see that the method is generally successful in identifying links. This is the case, even as the accuracy seen in Figure 6-6 has room for improvement.

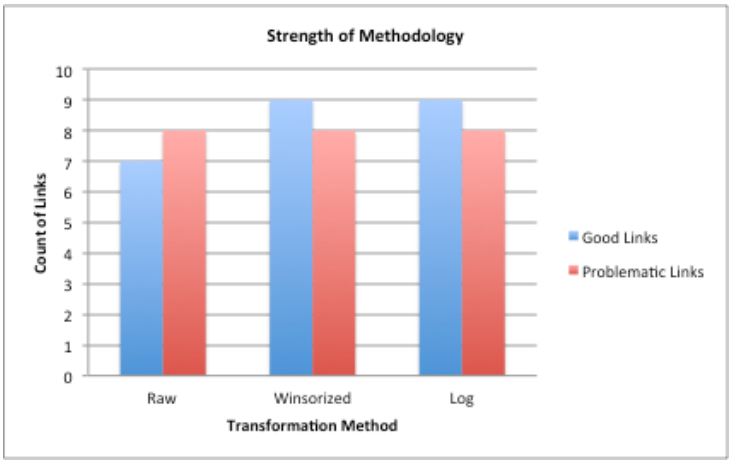


Figure 6-6 Strength of Methodology. This graph shows counts of the good and bad links correctly identified by the method grouped by data transformation method.

6.5 Discussion of Results

Prerequisite skill structures in any knowledge domain are very important for instruction and for preparing students for future learning. Almost every knowledge domain has one or a couple, which are created by domain experts. It is important to note that several of these prerequisite skill structures need to be refined. Data is currently being generated that affords us the opportunity to use data-centered methods to refine these structures.

In this study, we used data generated from PLACEments, which is an adaptive testing feature of ASSISTments, to propose one method for refining these structures. In this current study, we used a simple linear regression method in which we use the student performance on the prerequisite skill to predict their performance on the post-requisite skill. We found that for some of the links, this method was effective at identifying both good and bad links in the structure. Comparing the results of the method with the survey provided a ground truth with which to compare the findings from the study. The results have shown that if we have performance data, in the form of mastery speeds, we can achieve more accurate results by transforming the dataset in some format in order to take care of outliers that can easily ruin the findings. [61] The methodology affords prerequisite skill structure creators i.e., domain experts, the opportunity to identify and refine the order of the skills in these structures.

It must however be noted though that the method was not perfect. A few of the links could not be correctly identified. Additionally, the criterion for determining whether a regression model is worth examining is relatively low. In view of these, further studies are required to ascertain the reasons behind that finding and to propose refinements of the method. There could be interaction effects and other relevant predictors that have been ignored, but which may be necessary to ensure better accuracy for the prediction models.

6.6 **Contribution**

The main contribution of this study is the provision of a simple and effective linear regression-based method for refining prerequisite skill structures. With this method, we are able to identify problematic arcs in the structure and make these findings available to domain experts who will then use this information to further refine the maps. Additionally, we have introduced a system that provides us with a very good source of data for refining prerequisite skill structures.

6.7 **Conclusions**

Several authors in the educational data mining and learning analytics community have attempted to learn prerequisite structures from students' response data. Others have looked for methods to refine already existing, domain-expert-made prerequisite skill structures using methods like LFA, etc. It has so far been difficult to get datasets that present students' response data in the order of their underlying students' performance. This chapter uses such a dataset available from PLACEments, an adaptive testing system that traverses a prerequisite skill structure for item selection. The data from students' performance on remediation assignments was used to learn the strength of prerequisite skill relationships existing between skills and to make suggestions regarding these arcs. We have shown that using simple linear regression and with the right dataset, we can relatively tell how strong the prerequisite skill relationship between two skills are and, based on that make suggestions, regarding which links domain experts may need to investigate and refine.

This study does have some limitations. The data used for this analysis has come solely from PLACEments. There are not many such adaptive testing systems that generate the kind of data we used in this study. It will be interesting to study another dataset that is generated in the same or similar format as PLACEments, in order to apply this simple linear regression method to make statements about the strength of prerequisite skill relationships. We view this study as a preliminary step in our goal of finding optimal prerequisite structures using PLACEments data.

7 Modelling Interactions across skills: A method to Construct and Compare Models Predicting the Existence of Prerequisite Skill Relationships

The incorporation of prerequisite skill structures into educational systems helps to identify the order in which concepts should be presented to students to optimize student achievement. Many skills have a causal relationship in which one skill must be presented before another, indicating a strong skill relationship. Knowing this relationship can help to predict student performance and identify prerequisite arches. Skill relationships, however, are not directly measurable; instead, the relationship can be estimated by observing differences of student performance across skills. Such methods of estimation, however, seem to lack a baseline model to compare their effectiveness. If two methods of estimating the existence of a relationship yield two different values, which is the more accurate result? In this work, we propose a baseline model that can be used to compare not only different methods of estimating the strength of skill relationships, but also can be used to observe which student-level features are most accurate in providing these estimates. Focusing on interactions of performance across skills, we use our method to construct models to predict the existence of five strongly-related and five simulated poorly-related skill pairs. Our method is able to evaluate several models that distinguish these differences with significant accuracy gains over a null model and provides the means to identify that interactions of student mastery provide the most significant contributions to these gains in our analysis.

This chapter is published at the following venue:

Botelho, A., **Adjei, S. A.** & Heffernan N. (2016) Modeling Interactions Across Skills: A Method to Construct and Compare Models Predicting the Existence of Skill Relationships In Tiffany Barnes, Min Chi and Mingyu Feng (eds.) *Proceedings of the 9th International Conference on Educational Data Mining* held in Raleigh, North Carolina, USA. — June 29 - July 2, 2016 pp292-297

7.1 Introduction

Many educational systems like ASSISTments [42] and Khan Academy already implement a prerequisite structure as a suggested ordering in which skills should be presented to students. These structures are often developed by domain experts and teachers in the field of study and are likely to hold ground-truth. It is clear, for example, that

relationships can be identified by observing skills at the problem-level; by viewing the steps required for students to complete each item, it can be known that any skills required to complete such problems can be considered prerequisites (for example, Multiplying Whole Numbers as a prerequisite to Greatest Common Factors as is used in our analysis). While this is true from a content perspective, it is also possible, and perhaps comparably useful, to describe general skill relationships in addition to prerequisite arches; general relationships could refer to skills requiring similar processes to complete rather than a relationship of content or may be similar in any other content-independent aspects. Such relationships, perhaps even themselves existing as a causal arch (expressed as a prerequisite), may not be found by domain experts due to their non-intuitive relationship. Therefore, by using existing prerequisite links that are known to be strong by domain experts, we are able to construct a method of measuring which factors are most predictive of such relationships.

We also argue that identifying strong relationships is not enough for a method of prediction to be considered adequate. Such a method should also be able to identify weak or non-existent skill relationships. It is likely that while much attention and research is placed on structuring prerequisite links, some of the deemed strong links are false-positives. In other words, a skill may be listed as a prerequisite, but has no true relationship to its supposed post-requisite skill. In such a case, there is little or no interactions of performance. Such links must also be identified and removed or reordered in learning platforms to benefit the students.

A significant amount of research has looked at measuring the strength of skill relationships [1, 16], and even the effects such relationships have on measuring student performance [14, 93], but without a common baseline model, it is difficult to compare the true accuracy of such methods in a general setting. Furthermore, many of these methods represent similar conceptualizations of performance inherently, or through variations of representation such as aggregation or centering. For example, "student achievement" is likely a predictor of skill relationships (achievement on a prerequisite skill will likely influence achievement on a post-requisite skill), but can be represented as the percent of problems answered correctly, mastery speed (the number of items needed to complete an assignment as is commonly used in intelligent tutoring systems), or countless other combinations of features. It will be important to distinguish between these generalized components to avoid incorporating features that capture the same types of conceptualizations into predictive models.

This study attempts to provide a baseline model that can be used to compare and identify which features best indicate a strong relationship between two skills. This baseline model will incorporate a method of generalizing and distinguishing features that measure different aspects of learning and performance. With this model, we seek to answer the following two research questions:

1. *What link-level features, expressed in this* chapter *as interactions of performance between skills, are significant in predicting the existence or non-existence of skill relationships?*
2. *Are we able to identify which features are the strongest predictors of skill relationships, and if so, does combining them make for a more accurate predictive model?*

The next section of this chapter will discuss some of the previous research performed on skill relationships and prerequisite structures. Then, we will discuss our theory and methodology to provide a baseline model of comparing methods of measuring skill relationships. Using this model, we then compare several commonly-used student-level features, and of the most accurate, compare several different representations of those features. Finally, we will discuss our findings and suggested future works.

7.2 **Previous Work**

The discovery and refinement of prerequisite skill structures has been an important research question in recent years. The impact of this research on educational systems cannot be overemphasized. Domain experts who design these structures need data centered methods to support the decisions they make; it is vital to have empirical data to support hypothesis regarding the order in which skills are presented as it can have a large impact on student achievement and either aid or impede the learning process. Additionally, identifying the best prerequisite skill structure will enhance student modeling; knowing a student's prior performance on prerequisite skills can help estimate that student's performance on the post-requisites. This can lead to earlier interventions for struggling students, or even help redefine mastery perhaps students who perform very well on a prerequisite requires less practice on a post-requisite, or can be given more advanced examples.

Tatsuoka, defined a data structure called the Q-Matrix, that represents the mapping of problems to skills: the rows of this matrix represent the problems, and the columns represent the skills. [86] Though the goal of the research was to diagnose the misconceptions of students, they set in motion a number of studies that have used this data structure as the first step to find prerequisite structures [12, 21, 82].

Desmarais and his colleagues developed an algorithm that finds the prerequisite relationship between questions, or items, in students' response data. [31] They compare pairs of items in a test and determine any interactions existing between each pair. Depending on the interactions and a set of interaction-related criteria, they determine whether the two items have a prerequisite relationship between them. This approach was applied by Pavlik, et al. to analyze item-type covariances and to propose a hierarchical agglomerative clustering method to refine the tagging of items to skills. [64] Brunskel conducted a preliminary study in which they use students' noisy data to infer prerequisite structures [16]. Further research by Scheines, et al. extended a causal structure discovery algorithm in which an assumption regarding the purity of items is relaxed to reflect real data and to use that to infer prerequisite skill structure from data [82].

7.3 **Dataset**

The dataset[2] used for this study consists of real-world student data from the ASSISTments online learning platform. The raw data contains student problem logs pertaining to ten math skills from the 2014-2015 school year. These ten skills represent five skill pairs, listed in Table 7-1, for which domain experts identified as having a strong prerequisite relationship. While we are not limiting the usage of our proposed baseline model to just prerequisite relationships, these are the most reliable to identify due to the causal effect of content (if problems in skill B require the use of skill A to complete, a strong relationship can be identified)

[2] The full raw and filtered datasets are available at the following link: http://tiny.cc/veqg5x

Table 7-1 The strong skill pairs as determined by domain experts

Prerequisite	Post-requisite
Multiplication of Whole Numbers	Greatest Common Factor
Subtracting Integers	Order of Operations
Division of Whole Numbers	Dividing Multi-Digit Numbers
Volume of Rectangular Prisms Without Formula	Volume of Rectangular Prisms
Nets of 3D Figures	Surface Area of Rectangular Prisms

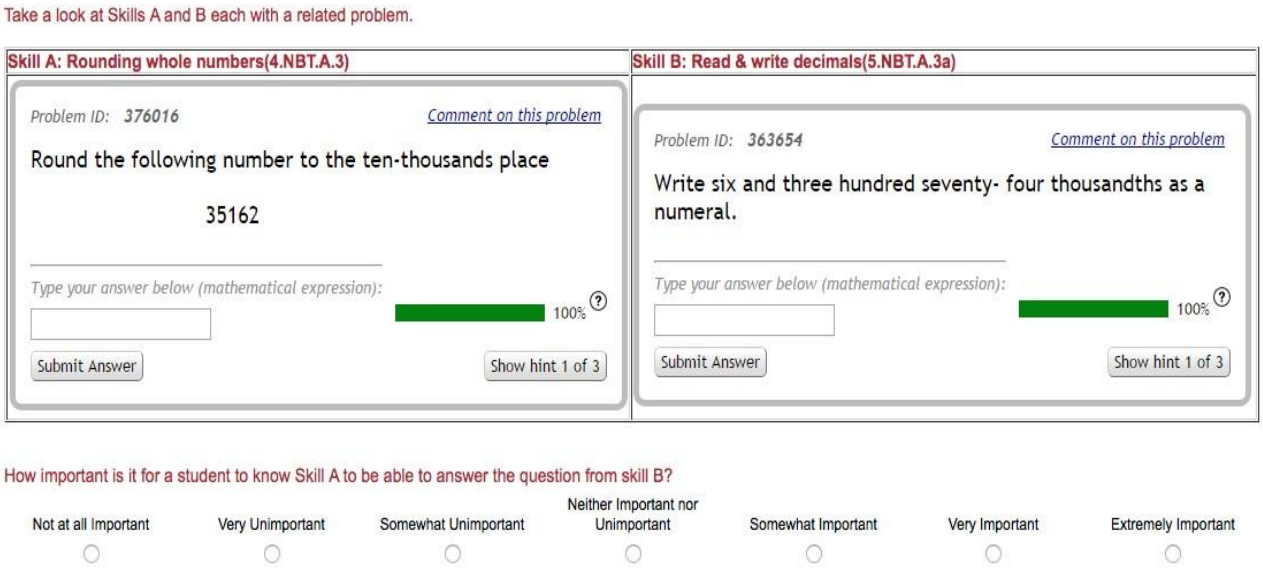


Figure 7-1 Sample question from the survey given to teachers and domain experts to help identify strong skill relationships

In order to identify believable ground-truth skill pairs, a survey containing 24 skill pairs for which we had sufficient student data (greater than 50 student rows) was administered to 45 teachers and domain experts who use ASSISTments. Each was asked to rate on a scale of 1 to 7, indicating the perceived qualitative strength of the relationship of each skill pair. A sample question from that survey can be seen in Figure 7-1. From the survey results, five skill pairs were selected to be the strongest related links with the smallest variance in opinion scores. As we are treating these links as ground truth, we wanted to be highly selective of these pairs.

The resulting dataset consists of 1838 student rows consisting of 896 unique students. This number of rows includes two rows of data per student for each of the five skill pairs included. The first row contains information of that student's

performance on the pre- and post-requisite skills, while the second row contains student performance on the prerequisite and a simulated post-requisite described further in the next section.

For each student, a feature vector was selected using common performance metrics to compare within our model. This feature vector contained eight link-level features representing the interactions between student-level prerequisite and post-requisite performance metrics. The generated link-level features observed are as described below:

Percent Correct: The mean-centered[3] percentage of correct responses in the prerequisite skill multiplied by the mean-centered percentage of correct responses in the post-requisite skill.

First Problem Correctness (FPC): The binary correctness of the first response in the prerequisite skill multiplied by the binary correctness of the first response on the post-requisite skill.

Mastery Speed: The mean-centered mastery speed of the prerequisite skill, defined as the number of problems required for each student to achieve three consecutive correct responses, multiplied by the mean-centered mastery speed of the post-requisite skill. In addition to centering, these values were also winsorized to make the largest possible value 10, chosen as this is often the maximum number of daily attempts allowed within ASSISTments. All centering and winsorizing occurred before multiplying the two values.

Z-Scored Percent Correct: The z-scored[4] value of mean-centered percentage of correct responses in the prerequisite skill multiplied by the z-scored value of mean-centered percentage of correct responses in the post-requisite skill.

Binned Mastery Speed (Bin): The numbered bin of mastery speed as described in [3] of the prerequisite skill multiplied by the bin of mastery speed in the second skill. Students were placed into one of five bins based on mastery speed if the assignment was completed and based on percent correct if the assignment was not completed.

Z-Scored Mastery Speed: The z-scored value of mean-centered, winsorized mastery speed in the prerequisite skill, multiplied by the z-scored value of mean-centered, winsorized mastery speed in the post-requisite skill.

---

[3] All centering of features was performed at the skill-level.
[4] All z-scoring was performed at the class-level.

Bin X FPC: The binned mastery speed value in the prerequisite skill multiplied by the binary correctness of the first response in the post-requisite skill.

Percent Correct X FPC: The mean-centered percentage of correct responses in the prerequisite skill multiplied by the binary correctness of the first response in the post-requisite skill.

7.4 **Methodology**

The ultimate goal of our model is to provide the means of comparing and identifying features that most accurately predict the existence (or non-existence) of skill relationships. As such, we propose a method that can be summarized in three parts. We develop a baseline model using known strongly related as well as simulated poorly-related skill pairs. Using principle component analysis, we group similar features into more generalized conceptualizations to both compare which types of features matter when predicting relationships, but also to avoid problems of multicollinearity that may bias our estimates. Once this baseline model is established, we can construct new predictive models from the significant features and observe their accuracy in predicting the existence of skill relationships when compared to a simple null, or unconditional model.

7.4.1 Baseline Model

We build a baseline model which we can use to compare combinations of features that estimate skill relationships. The theory behind this model stems from the idea of a "perfect" relationship-measuring method. Such a method needs to not only accurately identify strong skill relationships, but also must be accurate at identifying weak, or non-existent relationships. The dataset created from the five believable skill links will be used as ground truth strong skills, but we also need to introduce weak skill relationships into the dataset in order to ensure a more robust measure of accuracy.

In order to compare the usage of features against a weak or non-existent relationship, we simulated a new skill using students from the existing prerequisite skill by generating random sequences of responses. For each existing student, we randomly assign him/her a probability between 0.5 and 0.9 in order to create a random sequence of answers. For example, a student given a probability of 0.5 has a 50% chance of answering each given problem correctly. We simulate student answers until either mastery is achieved, defined as three sequentially

correct responses, or the student reaches 10 problems without mastering; a value of 10 is chosen here, as many assignments in ASSISTments are given a daily limit of 10 problem attempts. While we acknowledge that there are many ways to accomplish this simulation step, we feel this simple method sufficiently creates a skill that has no relationship to the original prerequisite as intended. As our proposed method is intended to be used in the future to help identify undiscovered pre- or post-requisite links, we chose to use a simulated skill rather than a random existing skill to avoid the possibility of randomly selecting an undiscovered relationship. Again, we wanted to be highly selective and consider several such scenarios as we are attempting to create ground-truth values to which we can make our comparisons.

Using these two skill-pairs, one link representing a strong relationship while the other representing a non-existent relationship, we can calculate a feature vector for each student in the prerequisite skill with values from each skill-pair. The purpose of this study is primarily to provide a method of comparison, and therefore choose features to observe based on commonly studied metrics, focusing primarily on feature interactions across skills. Future work could use our method to expand on this in order to look at other models and feature representations.

Using the existence of a relationship (either a relationship exists or it does not) as a dependent variable in a binary logistic regression, we compare the set of student features based on their predictive power. By identifying the features that best predict skill relationships, this model can later provide a value representative of the strength of each skill link.

Table 7-2 The results of the PCA analysis. All features except Z-Scored Mastery Speed mapped to one of three generalized components

	Component		
	1	2	3
Percent Correct		.821	
First Problem Correctness (FPC)			.839
Mastery Speed	.969		
Z-Scored Percent Correct		.865	
Binned Mastery Speed (Bin)	.972		
Z-Scored Mastery Speed			
Bin X FPC			.873
Percent Correct X FPC		.612	

We begin to compare commonly used student-level features in this study through two levels of experimental analysis. The first experiment performed with our proposed model attempts to compare groups of features, generalizing different representations of similar features into conceptual groupings. As such, we are able to view the predictive power of what we denote as initial performance, mastery, and correctness. The second experiment looks at the individual features as different representations of the overall group to compare these predictors at a closer level. We can take each factor of mastery, for example, and compare their usage in several models to determine which is the most accurate predictor of the existence of skill relationships.

7.4.2 Comparing Link-Level Features

In order to compare representations of student-level features, we must first be able to compare general conceptualizations of features to determine which provide more accurate predictions of the existence of skill relationships. We want to capture the true representations of each metric and attempt to interpret these generalizations as types of features. In order to accomplish this grouping of predictors, we use principal component analysis (PCA) to identify which student-level features correlate to and are representative of more generalized components. PCA is primarily used for dimensionality reduction as we are doing here and gives us the ability to create new variables from the component mappings. The resulting feature alignment can be seen in

Table 7-**2**. As is the case in our study, and was mentioned in the previous section, we have multiple metrics of mastery speed as well as several other features. As we can represent "mastery" in several ways, we want to know if the overall concept of mastery, as captured by the metrics used, is reliably predictive of the existence of skill relationships. The usage of PCA in comparing the groupings of similar predictors also helps to avoid problems of multicollinearity in our comparisons. As we are relying primarily on the significance values of each feature when predicting skill links, incorporating features that overlap in what they represent could bias results and alter such values. Multicollinearity is difficult to avoid entirely in this case, as many of the metrics used describe student general performance to some degree and are unlikely to be entirely independent of each other, but our method attempts to at least remove the larger effects of this problem on our results.

Creating a new set of predictors of these groupings, we are able to incorporate these into a binary logistic regression model to view the predictive power of each. While PCA groups similar features together based on their correlations, by viewing which features are grouped we are able to interpret and label each. From this process, we found that most of our features fell into three categories for which we have given the names "mastery," as this consists of representations of mastery speed, "correctness," as this consists of representations of the percentage of correct student responses, and "initial performance," as this consists of representations of student performance on the initial items of each skill. In addition to these three categories, we are also left with student mastery speed z-scored within student classes as a variable that did not fall under either of the three aforementioned categories; while a derivation of mastery speed, we believe that this did not correlate to the "mastery" category due to the method of standardization as it is capturing this metric in relation to students' peers. We will readdress this case in our section of discussion.

Once these predictors are identified and created, we construct a binary logistic regression model to predict, for each student row, whether a relationship exists or not. This model will give us a significance value and coefficient for each predictor in the model, as well as an overall predictive accuracy of the model which will be used more for the next analysis.

7.4.3 Comparing Feature Models

After being able to compare which generalized groups of features are significant predictors of the existence of skill relationships, we are able to compare the individual student level features that fall into each category by incorporating them into separate models to observe predictive accuracy. The analysis of the first experiment is used to determine which categories are significant in predicting the existence of skill relationships. Using that information, we are able to focus on those groupings with significance to construct models that utilize factors from each grouping. The grouping of "mastery," for example contains the factors of mastery speed and binned mastery speed, so we can construct models using each to compare differences in predictive power. To avoid problems of collinearity, no single model contains more than one factor from a single grouping. This significantly reduces the number of combinations of features to test compared to running this experiment without first grouping like features and identifying those that are significant as we did in the first experiment.

Using the significant groupings, we are able to create 17 models consisting of single, pairs, and triplets of features. A logistic regression is run on each of these models to predict the existence of a skill relationship. Of the 17 models, 10 of them produce a statistically significant prediction when compared to a null model. Ideally, our null model should produce a 50% accuracy as there is an equal number of good and bad link rows in our dataset. This is not always the case, however, as depending on the feature observed, information may be missing for a particular student; mastery speed, for example, as the number of items attempted by a student before reaching 3 consecutive correct answers, would be missing for any student that did not complete the assignment. For this reason, the predictive power of each model is described as gains in predictive accuracy, or rather, the accuracy of each model minus the accuracy of the corresponding null model. In the case of our analysis, the null model ranges in accuracy from 50% to about 62% depending on the factors contained within that model.

Table 7-3 The coefficients and significance values of the generalized components analyzed. From this we can focus on models that exclude features contained in the components with no significance.

Component	Coefficient Value (log-odds units)	Significance
Mastery	-.251	<.001
Correctness	.015	.802
Initial	.129	.037*

Performance		
Z-Scored Mastery Speed	-.129	<.001

7.5 **Results**

The results of the first experiment are expressed in Table 7-3. Each of the three feature groupings of Mastery, Correctness, and Initial Performance created using PCA in addition to the Z-Scored Mastery are compared within the same model, predicting the existence of a skill relationship. As these again are link-level features describing interactions between student-level performance on prerequisite and post-requisite skills, it is difficult to draw tangible interpretations from the coefficient value, expressed in log-odds units. This coefficient, used in the logistic regression to make the predictions, describes each component's effect on the dependent variable. For example, for each unit increase in "Mastery," the probability that the link exists decreases. Again, as this component is an aggregation of interaction features, it is really describing an aggregation of differences of differences between student-level features making it difficult to make definitive claims regarding these values alone and were included purely to display a general trend of these components on the prediction.

From Table 7-4, we see the significance of each component on the overall prediction by viewing the corresponding p-values in the fourth column. Looking at these values, we can claim that the overall grouping of "Correctness" seems to have less of an impact on the predictive accuracy of the model. As this term is not significant, we can focus the remainder of our study on the remaining three components.

Figure 7-2 illustrates the results of our second analysis comparing the models that we are able to construct with the remaining features once the "Correctness" grouping has been disregarded. This figure shows the comparative predictive accuracy of the 10 models that give statistically significant predictions as seen in Table 7-4. Again, these values are expressed as accuracy gains, or rather the percent accuracy increase over the null model run for each predictive model.

Table 7-4 The models constructed from features in the significant generalized components. No one model contains more than a single feature from each generalized component.

Model	Null	Model	Accuracy	Significance

	Accuracy	Accuracy	Gain	
Mastery Speed (MS)	0.63	0.62	0.00	1.000
Z-Scored Mastery Speed	0.63	0.63	0.00	0.888
First Problem Correctness (FPC)	0.50	0.56	0.06	<0.001***
Binned MS	0.50	0.69	0.19	<0.001***
Bin X FPC	0.50	0.56	0.06	<0.001***
Bin, Z-Scored MS	0.50	0.71	0.21	<0.001***
MS, FPC	0.63	0.62	0.00	1.000
MS, Bin X FPC	0.63	0.62	0.00	1.000
Bin, FPC	0.50	0.69	0.19	<0.001***
Bin, Bin X FPC	0.50	0.69	0.19	<0.001***
MS, FPC, Z-Scored MS	0.63	0.63	0.00	0.754
MS, Bin X FPC, Z-Scored MS	0.63	0.63	0.00	0.979
Bin, FPC, Z-Scored MS	0.50	0.71	0.20	<0.001***
Bin, Bin X FPC, Z-Scored MS	0.50	0.71	0.21	<0.001***
MS, Z-Scored MS	0.63	0.63	0.00	0.843
FPC, Z-Scored MS	0.50	0.64	0.14	<0.001***
Bin X FPC, Z-Scored MS	0.50	0.61	0.11	<0.001***

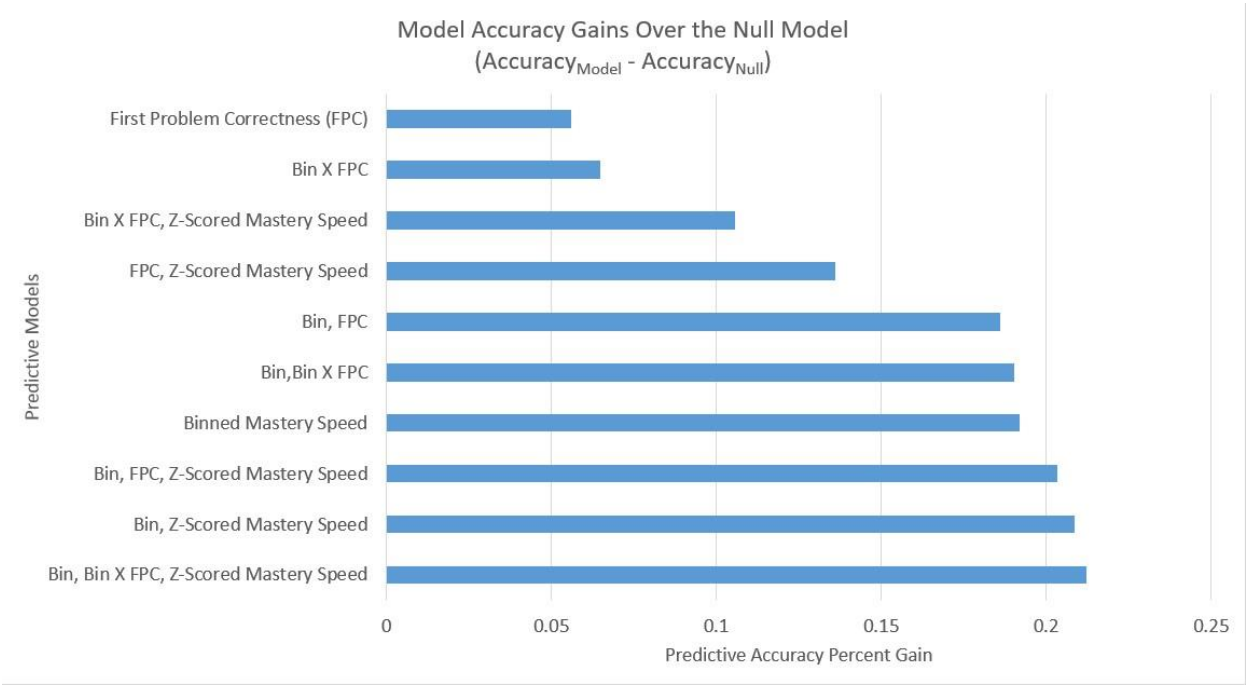


Figure 7-2 A comparison of accuracy gain of the models with statistically significant predictions. The Bin feature is found in the most accurate models indicating that it is a strong predictor.

7.6 Discussion

This chapter provides a baseline model of comparing student level performance across skills to measure the strength of a skill relationship and compare the accuracy of both features and models that estimate this value.

Such a model, in our experience, has not existed prior to this study. Our method attempts to identify not only the individual features that contribute to better predictions of these relationships, but also moves to generalize similar features into conceptualizations for comparison in order to minimize multicollinearity.

The principle component analysis step of our model found that all but one feature mapped to one of three components that we have interpreted as mastery, correctness, and initial performance. It was found the z-scored mastery speed, contrary to our intuition, did not map well to the grouping of mastery. We can speculate the reason for this occurrence by altering our interpretation of the feature. Mastery speed itself is an interesting metric as it attempts to capture two dimensions of performance: a level of understanding and a rate of learning. Also, to reiterate a prior distinction, these metrics are interactions of performance across skills. When centering, winsorizing, and z-scoring this metric, it seems to have changed its representation and we can alter our interpretation of this feature to be the change of comparative performance to one's peers across skills.

Observing the resulting model components from the principle component analysis in Table 7-2, we were able to focus our attention to those components with significant values. Correctness was the only component of that model that was found to have no statistical significance on the dependent variable. This is certainly interesting, as percent correctness and other such measures are among the most common metrics of performance. Perhaps the interaction between pre- and post-requisite percent correct is losing some predictive power from when the metric is used for other predictions of performance.

This aspect illustrates one other important finding that the distinct representations of one metric or another each contribute differently to the predictive accuracy of the models studied. Models incorporating mastery speed, for example, had no significant accuracy gains over a null model, while mastery speed binning showed considerable gains as seen in Table 7-4. The baseline model of comparison proposed in this study provides the means to make that distinction regarding features contained within the same generalized component grouping. As is seen in that figure, combinations of features outperform any single feature, illustrating a more robust model by capturing multiple representations of performance.

7.7 Future Work

While we have shown that our model is able to compare and identify features that contribute to higher accuracy in predicting the existence of skill relationships, we also need to stress the importance of the usage of this information. The ability to compare features is only the first step of our model's goal. By identifying strong predictors of skill relationships that we know exist, we can apply it to other skills within ASSISTments and other systems to identify potentially new prerequisite arches, and also to better measure and predict long-term student performance, learning, and retention. Having an accurate estimate of skill relationships can help restructure prerequisite structures to provide skill sequences in an order that optimizes student learning and achievement.

The work in this chapter incorporated several skills into a single dataset to make predictions. In this case, we wanted to create a method that is generalizable to some degree. While our selective skill set allows us to make some claims in terms of the accuracy these models over all skills, it may likely be the case that skill relationships are measurable in different ways for different skills. Further analysis could repeat the steps here on each one of the acquired skills in the dataset. While correctness was not significant in these results, perhaps it is significant when predicting certain types of skills. Perhaps, similar to our features, skills themselves could be generalized into conceptual types for different kinds of analysis pertaining to interactions of performance and their relationships.

The feature vectors generated for each student in our dataset captured many of the most common student-level metrics, but certainly not all of them. There are many other aspects that could be added including completion, measures of learning rate, time spent on the assignments, hint usage, and countless other variables. In addition, this study only observed interactions expressed as multiplications of these terms to describe them as link-level features. There are various other ways to represent interactions or other such transformations including differences of values, division of values, or just simply cross-feature interactions as was partially explored here by looking at Bin X FPC and Percent Correct X FPC. Such interactions model various other aspects of student performance and behavior that can be very useful in this type of relationship prediction.

In addition to all these, this model could potentially benefit from personalization by adding a measure of student prior performance before even the prerequisite. Perhaps a skill is measurable to different degrees depending on the type of student in terms of past data.

8 A Correlation-Based Method for Inferring Prerequisite Skill Links from Learner Performance Data

Using learner performance, rather than domain expertise, to infer prerequisite skill links has been a topic of recent interest. Several different data mining methods have been proposed to infer prerequisite links. In this chapter, we present preliminary results from a method that uses partial correlations to infer these links from learner performance data. Our method starts out by generating a number of links and pruning the generated links based on a series of criteria. These criteria are applied sequentially to the intermediate results of the method. The pruned list of links is then presented to domain experts for their judgements. We find that our method infers meaningful prerequisite links. After conducting a survey of human experts on a subset of the links inferred by our method and domain-expert designed links, we find that there was higher rate of agreement (80%) among the experts on the data mined links, than even their own links (70%). The pruning method presented herein can be used to augment other data mining methods, in order to reduce the number of spurious links that they generate.

8.1 Introduction

In recent times, there has been a proliferation of studies that investigate methods for inferring prerequisite skills from learner performance data. Many of these methods can be categorized into two main groups: 1) pure data mining-based methods, and 1) experiment-based methods. Most of the research in this area apply to the use of pure data mining methods for refining prerequisite skill graphs.

The many pure data mining methods that attempt to identify latent skill structures address the question of whether or not the current skill topologies represent the best sequencing of content for instruction. Can we improve upon the models by identifying missed relationships or removing weak or non-existent links? In fact, the question of learning prerequisite skill structures from learner performance data is beginning to gain traction in the educational data mining and analytics community in recent years. Among the numerous methods that have been proposed for this exercise is Partially Ordered Knowledge Structures, which was proposed by Desmarais et al. [28]. POKS generates item-level topologies based on the correlation between student performances on items or problems. Learning Factors Analysis (LFA) was introduced as another method for refining existing prerequisite

skill topologies, and possibly generating new topologies. [65] Pavlik et. al. extended the LFA model in their quest to generate domain models, also referred to as skill topologies. [63] They analyzed learning curves and used the results of their analyses to generate the learning models from student performance data. Additionally, Chaplot, et. al. combined text-based and performance-based methods for inferring prerequisite skill topologies from student performance data as well as from text-based course materials. [20]They proposed an unsupervised approach for the task, and report that this approach outperformed alternative supervised methods. [21] Chen and colleagues proposed a probabilistic association rules mining method for discovering prerequisite skill structures from learner performance data. What makes their method different from the others is the use of a learned evidence model. [21] The model is used to estimate students' probabilistic knowledge states as they progress through an assignment.

Adjei and colleagues have used randomized controlled trials (RCT) [1] as well as predictive models [3] in PLACEments to infer prerequisite links between skills. PLACEments is an adaptive testing and remediation system that is based on a prerequisite skill structure. Both studies identified sections of the prerequisite structure that need improvement. The sections that require improvement have issues that include: 1) links that are in the wrong direction, and 2) links that have prerequisites which do not help students to demonstrate competence in the post-requisite skill. The problem with using RCT's in PLACEments for this purpose is that, it requires a lot of time to collect sufficient data to be able to draw meaningful conclusions from the experiments.

In this chapter, we describe a correlation-based method for inferring prerequisite skill links between knowledge components based on learner performance data and inputs from domain experts. The chapter is organized as follows: the next section describes the methodology. We then present the results and a comparison with domain-expert designed links. Finally, we conclude with a discussion of the benefits and limitations of this approach as well as planned future work.

8.2 **Methodology**

This section describes the dataset for the study, the partial correlation-based method for generating prerequisite skill links and the pruning and evaluation method.

8.2.1 Dataset

The data set used for this study is pull from ASSISTments [42], a predominantly math-based homework and tutoring system. This system is known to cause learning gains for students using the system [79]. ASSISTments provides students with problem sets. There are different types of problem sets, the predominantly used type being Skill Builders. Each such problem set is called a skill builder. Each skill builder examines students understanding of a specific math skill (aka, knowledge component). A student is said to have mastered a skill builder/knowledge component if he/she answers 3 items correctly in a row without being given any assistance. As long as that streak of correct answers is not achieved, the system continues to provide more questions until the student reaches a daily limit of questions (which is usually 10). When the daily limit is reached without the student reaching mastery, the assignment is discontinued until the next day. The idea behind the daily limit is to force students to seek further assistance from elsewhere. This is due to the understanding that the assistance provided on the ASSISTments platform for that skill builder has not been helpful to the student. The number of items it takes the students to complete a skill builder is referred to as the *mastery speed* of the student for that skill builder.

In this data set, if a student could not complete the assignment at the time of data collection, the student's assignment is given a mastery speed of 20. This was to minimize data loss. Additionally, we felt that students who dropped out from an assignment will give us useful information about the particular skill being tested in the assignment. The higher the mastery speed, the worse the student's performance. This will result in higher skill penalty for our method. The dataset contains skill builder assignment performance data for 22,023 8th grade students. The students were assigned at least two of the 120 different skill builders in the data set. Some students were assigned the same skill builder multiple times. For those students, we included their performance on the very first attempt on that skill builder. This was done in order to eliminate data pollution that may be caused by over representation of performance data for the same skill builder-student pair.

8.2.2 Generating and Evaluating Skill Links

Generating skill links is performed in two phases: 1) Identify all possible links and 2) prune the identified links based on a set of criteria. The first phase is purely data mining based whereas the second phase is where expert input is required to ensure that the identified links make sense. This section describes both phases in detail.

8.2.2.1 Generating Possible links: The Partial Correlation Based (PCB) Method

We generated ${}^{n}P_{2}$ - n different permutations of the n different skills. To illustrate this, suppose there are 3 different skills (A, B, C) in our dataset. This results in 9 possible links (shown in Table 8-1)

Table 8-1 Possible Permutations of skills in from a set of 3 skills (A, B, C)

Row #	Prerequisite Skill	Post-requisite skill
1	A	A
2	A	B
3	A	C
4	B	A
5	B	B
6	B	C
7	C	A
8	C	B
9	C	C

We removed all links resulting from skills that are linked to themselves, hence the "- n" term. In Table 8-1 for example, rows 1, 5 and 9 will be removed from the list of possible links to be investigated, resulting in 6 possible links. For each pair of skills that we generated, we identified students who were assigned skill builders for the two skills in the pair and in the order of the pair. To be more specific, suppose the pair <Skill A, Skill B> is one of the possible permutations of skills in the dataset (row 2 in Table 8-1). For this permutation, students get assigned Skill A and then Skill B after completion of Skill A. We therefore are interested in determining whether the relationship Skill A → Skill B is a valid prerequisite relationship that can be inferred from the data.

With this link, we identify students who were assigned and started the skill builder assignments in the order of the link. We then determine the correlation between the mastery speeds of all the students who started

both skill builders, partialling on their average mastery speed on all other unrelated skill builder assignments. The reason for partialling out the average mastery speed on all other unrelated skill builder assignments is to ensure that we remove the effect of the student's performance in general and to account only for the pure relationship between the two skills. The magnitude and sign of the partial correlation are then used as a measure of the strength of the relationship between the two skills in the link and the directionality of the relationship, respectively.

8.2.2.2 Pruning generated links:

While partial correlations are a good starting point, it leaves unspecified which specific links to focus on, and which spurious links to ignore. There are several reasons for the numerous spurious links. This section describes some of the reasons for the spurious links and how we pruning those links from our final result.

Spurious links: unreasonable directions. Since the permutations are agnostic to the grade levels of the individual skills that are paired up in a link, several spurious links can be learned. For example, you might find situations in which a second-grade skill is identified as a post-requisite to a 6th grade skill. This orientation is not reasonable, even though there could be data showing a teacher assigned the skill builders in that order. A possible reason for this is that teachers may want to assign lower grade skills to a subset of students because these students may be lacking those lower-grade skills, or some students struggled on the initial assignment and the teacher later assigned a prerequisite. We therefore applied the criterion to remove any links in which lower level grade skills are post-requisites of higher level grade skills.

Where to focus*: *strong partial correlations. We believe that for a link to make any sense and be included in the learned prerequisite skill graph, it must have a strength above a certain threshold. We set our threshold to 0.15 because relatively few links were that strong and so we screened out most of the links with very low partial correlation numbers. Links with higher values than the threshold were then included in the final inferred skill graph.

Spurious link*s: *non-similar sub-topics. Though this is not a recommended pruning criterion since it involves domain expert input, it reduces the number of links that could be found in the dataset and which can be analyzed by human domain experts. While our method is intended to use data to infer prerequisite skills, the method needs

to be informed by the static information about the domain. Examples of such domain information are the grades in which the skills are taught and the sub-areas to which the skills belong. In other words, the method capitalizes on what is already known about the knowledge components.

The number of links resulting from applying the above-mentioned criteria was large. We therefore decided to restrict the links learned to skills within the same sub-topic of the domain, applying static domain knowledge. In this context, a sub-topic will be similar to a strand in the Common State standards for mathematics [17]. Typical examples of sub-topics are "Expressions and Equations" and "Number Sense". In other words, two skills will be accepted to be in a skill link if they are both within the same sub-topic. It must be noted that this criterion eliminates the number of cross-topic prerequisite links that our method can learn. However, applying it ensures that human experts can have a reasonable set of links to study and make sense of. Additional studies will be conducted to consider the other cross-topic links that were identified.

8.2.3 Comparison with Domain Expert - designed Links

We compared a random selection of data mined links with domain expert designed prerequisite links and present the results of that comparison. We further presented 10 random selection of the links to three domain experts. These experts were asked to rate the goodness of the data mined links.

8.3 **Results**

Applying PCB method to the dataset described above, we identify a large number of prerequisite skill links, as was expected. Of the 120 skills in the ASSISTments Data set, we started out with 14280 possible links. There was a total of 2926 links for which were had student performance data. Applying the pruning criteria described in the methods section, reduced the number of links from 2926 to 56. As Table 8-2 and Figure 8-1 show, of the 178 links inferred, there were 56 links that were found to be within the same strand or topic area. By comparison, the graph constructed by a domain expert contained 156 links. This result of automated techniques finding fewer links than humans is a common result [12, 28]. PCB and the human-generated prerequisite graph (HUM) agreed about only 12 links. This accuracy is rather low, if we were to accept the HUM prerequisite skill graph as ground truth. Furthermore, the PCB method learned 3 links that were reversed in the domain expert designed graphs.

Table 8-2 The results of the application of Pruning Criteria

Category	Number of Links
Potential Links	14,280 (= $^{120}P_2 - 120$)
Links for which we had data	2926
Links with correlation > 0.15	304
Links with correct direction	178
Links in the same sub area	56

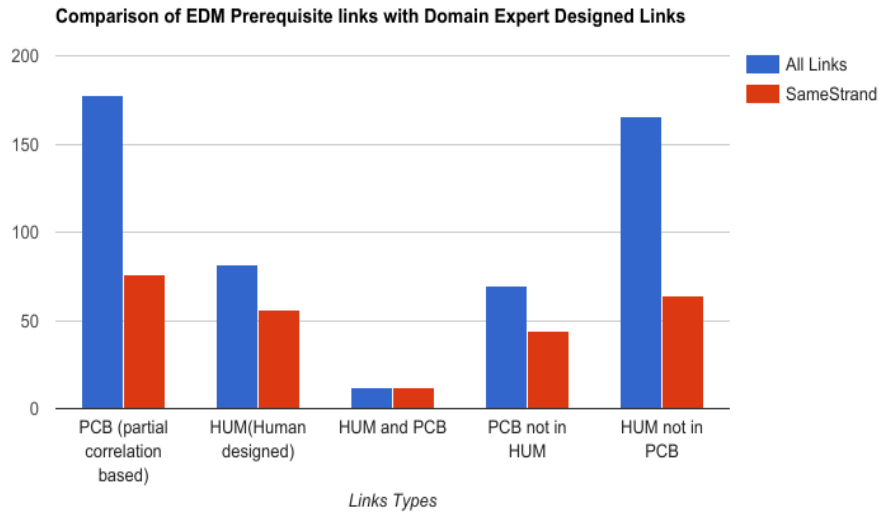


Figure 8-1 Number of links inferred using PCB method compared with Human designed links. The red bars show the links that are within the same strand and the blue bars are links in the correct direction.

Figure 8-2 shows the level of agreement between the domain experts. A large percentage of the randomly chosen skill links from the PCB links, which are not in the HUM links, were agreed upon to be worthwhile and reasonable links. This percentage is higher than percentage of links in the HUM links not present in the PCB links. This is quite surprising because it shows that the domain experts agree a lot more with the data mined links than with their own links. This gives credence to the usefulness of the PCB method.

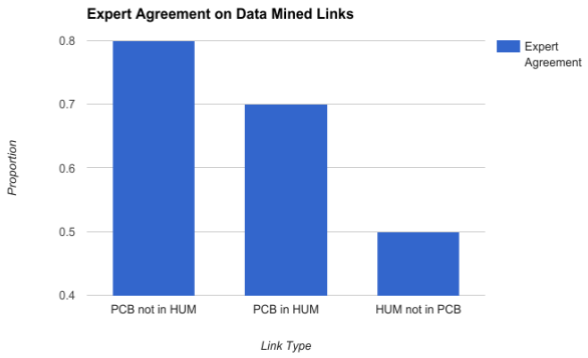


Figure 8-2 Domain Expert Survey Results

8.3.1 Survey of domain experts

We randomly selected 10 links each from the following three groups of links: Links that are both in PCB inferred links and domain expert designed links (i.e. "PCB in HUM"), domain expert designed links not in PCB mined links ("HUM not in PCB") and PCB inferred links that are not in the domain expert designed links ("PCB not in HUM"). Two domain experts who designed the HUM skill graph were then asked to rate each of the 30 randomly selected links. The experts were agnostic to the source of each of the 30 links. These experts were asked to indicate whether or not the links learned were reasonable. Figure 8-2 indicates the distribution of the responses from the experts. Each bar represents the proportion of links the experts agreed were reasonable. The experts agreed that 80% of the links found from the PCB links were accurate as compared with 70% from both the PCB and HUM links. It was surprising to note that the experts found only a small percentage of links from the HUM links to be reasonable.

8.4 **Contribution**

The major contribution of this work is the introduction of an approach for pruning prerequisite skill links from data mined skill graphs. The partial correlation-based pruning approach presented in this work ensures that we measure the true relationship between two skills. Partialling out the students' performance on other related skills eliminates the effect of the students' ability level on the existence of a relationship between two skills. Other data mining techniques for inferring prerequisite skill graphs could be augmented with this method.

8.5 Future Work

Our work has a number of limitations. One such limitation is the fact that we eliminate many cross-topic links. Additional work is required to investigate these other types of links that we pruned out of the data-mined graph. There could be a lot of meaningful findings that the pruning techniques ignore in our quest to reduce the size of the eventual graph that is learned. There is the need for a better method to eliminate spurious links from the data mined graph. Another important source of information that could be helpful in refining our search method is the inclusion of student help seeking behavior during assignment tasks. (whether they ask for help when practicing skill related tasks).

In our work, we compared the results of the data mined links with those of the human experts. We looked at the similarities and differences between the data mined links and expert designed links. As the results show, there were a few links that were completely different from those designed by domain experts. The next natural step that follows from this is to test which of the links better improves student learning. We intend to run randomized controlled trials to compare the data-mined links with the expert designed links. The results of these studies should inform domain experts regarding the prerequisite skill links to consider when designing these models.

8.6 Conclusions

We have attempted to infer prerequisite skill links from student performance data, just by using correlations between student performance on skills, partialling out their performance on some other non-related skills. We have also introduced a method for pruning spurious links. We compared the data mined links to domain expert designed links based on the same set of skills used in the PCB method. We also conducted a survey of domain experts and found that these designers may be missing useful links in the work. This preliminary result is an indication that using simple methods like partial correlations can be helpful in augmenting the other methods of inferring skill graphs from student performance data.

9 Improving Learning Maps using An Adaptive Testing System: PLACEments

Several efforts have been put forth in finding algorithms for identifying optimal learning maps for a given cognitive domain. A few authors have used Learning Factors Analysis and Q-Matrix-based algorithms for improving the predictive powers of learning maps with a degree of success. In [1], we proposed a greedy search algorithm for searching data fitting models with equally accurate predictive power as the original skill graph but with fewer nodes/skills in the graph. In another unpublished work, we showed that the algorithm is susceptible to a number of factors including the number of students, items as well as the initial guess and slip rates of the individual skills in the graph. In this chapter we present PLACEments, an adaptive testing system, and report on how this is used to determine the strength of the prerequisite skill relationships in a given skill graph. We also present preliminary results that show that different learning maps need to be designed for students with different knowledge levels.

This chapter is published at the following venue:

Adjei, S. and Heffernan N. (2015). Improving Learning Maps Using an Adaptive Testing System: PLACEments. *Artificial Intelligence in Education*. C. Conati, N. Heffernan, A. Mitrovic and M. F. Verdejo, Springer International Publishing; 2015-01-01. 9112: 517-520.

9.1 Introduction

In order to improve upon student learning, a number of approaches have been studied to determine ways of improving the effectiveness of the skills teachers transfer to students and the order in which these skills need to be taught. A few other methods have been proposed and used to improve upon learning maps. Cen, Koedinger, and Junker described a process for analyzing multi-dimensional skill maps whereby successive adjustments to a map were analyzed to determine the arrangement of nodes and connections that best fit available data. [19] Desmarais, et. al. present a framework for identifying structures from students' data and call these structures Partial Order Knowledge Structures (POKS). [28]The POKS framework is probabilistic in nature and infers the structure from the student's responses to a poll of items. One issue with this approach is its use of items. When the number of

items is large, the approach is not scalable. These and many other approaches have shown promising results, though many have not been applied to different subject domains. Wang proposed a genetic algorithm based method for determining optimal curriculum for schools. [95] The method proposed in Wang's study significantly reduced the amount of time needed to arrange the optimal curriculum, the level of granularity was at the course level and not at the level of individual knowledge components (skills in the learning map).

Given student responses to a number of questions, a prerequisite skill graph, in the form of a Bayesian Network, can be identified. Friedman and his colleagues proposed an algorithm they call the "Sparse Candidate" algorithm for learning such belief structures. Their method was found to be faster because it uses a heuristic to reduce the search space. [39]

While these approaches have their strengths and weaknesses, none of the approaches present a method by which to determine the strength of the relationships in a prerequisite skill graph. In our quest to find the best methods for improving the predictive and representative powers of learning maps, we observe that one possible method of improving skill graphs is to determine the strength of some of the links in the graph and to propose changes to the graph based on those strengths. Could we use empirical studies to determine the strength of the relationships between skills in learning maps and hence to determine whether these links belong in the graph? This chapter presents an adaptive testing system that traverses a prerequisite skill graph based on a student's performance. We present a brief description of how the system works, the design of the study, our method and the results we found. We also include an analysis of the results based on the knowledge level of students. The chapter concludes with a discussion of the findings of the study as well as the limitations of this approach.

9.1.1 PLACEments, an Adaptive Testing Systems

PLACEments is a computer aided adaptive testing system. This system is a feature of an intelligent tutoring system, ASSISTments, that mainly provides teachers a means of creating and as-signing exercises and tests to their students. [42] ASSISTments also has a feature that allows students to do the exercises and tests assigned by their teachers.

The PLACEments system has a number of components: item pool, item selection, termination rules (a skill graph for the knowledge domain in which the students' knowledge will be assessed and three modules: test creation, test taking (Tutor), and remediation creation). The item pool for placements is chosen from a list of skill builders used extensively in ASSISTments. The choice of problems for the placements test was made based on the difficulty of the item. The difficulty was determined by calculating the percent correct for all responses of students in ASSISTments to that item and subtracting that value from 1. (See equation 1 below) To ensure that PLACEments does not present overly easy or overly difficult problems, the items for each of the skills tested were chosen such that their difficulty is between .4 and .6 (the smaller the number, the more difficult the item is).

$$difficulty_i = 1 - Pr(item_i = 1) \qquad (1)$$

As noted earlier, PLACEments uses a predefined skill graph to guide test item selection. Though we currently use a prerequisite skill graph developed based on the Massachusetts Common Core State Standards for Mathematics [17] the system is designed such that it can use any prerequisite skill graph from which tests can be drawn. The initial set of problems is chosen from the initial set of skills chosen by the assigner of the test. Each skill has one problem chosen from the item pool. When students get an item for a skill incorrect, implying that the student does not have that cognitive skill, the test is expanded by including the problems from the prerequisite skills of the skill the student has gotten incorrect. The test bank increases until the grade boundaries chosen at test creation are reached. For a given student, the test terminates when all the skills in the initial set of skills have been tested, and the student gets all the items for that skill correct. If the student is not able to answer any of the initial problems correctly, then the test terminates when there are no prerequisites remaining to be tested.

The following diagram, Figure 9-1, shows a hypothetical graph that explains how the test proceeds. The correctness indicator attached to each node in the graph is a particular representation of a given student's performance. The nodes in the graph represents the skills, the arrows between the skills represents the prerequisite relationship between the skills (thus, skill 'D' is one of the prerequisites of skill 'A'). In this configuration, the students are assigned skills 'A', 'B' and 'C' as the initial skills. This student is adaptively assigned questions D and E since he gets the question for "A" incorrect. The student also does poorly on E and hence is presented a problem from H. The test terminates when the student reaches skill H since it is the last in the boundaries

specified at the test creation. The size of the test is affected by the students' performance as well as the structure of the skill graph.

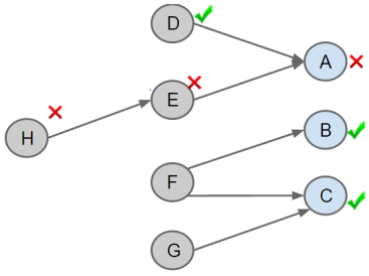


Figure 9-1 A sample skill graph and a sample student's response configuration

After a student completes the test, remediation assignments are created based on the skills they performed poorly on. Each student is assigned a different set of remediation assignments. The remediation assignments on the lowest grade level skills are released before those of the higher-grade levels. In the example shown in fig 18, the student will be assigned a skill builder assignment in the following sequence: H, E and then A. Once the prerequisite skill is completed, the next skill-related assignment inline is released.

9.1.2 Research Question

As was stated earlier, this study is meant to determine whether a skill graph can be improved using empirical studies. To be specific we want to determine the strength of prerequisite skill relationships between skills and hence determine which of such relationships to remove or maintain in a skill graph.

9.2 **Methodology**

To answer our research question, we run a study in which the navigation of a skill graph in a series of PLACEments tests is modified for a random sample of students. Figure 9-2 demonstrates the modifications made to PLACEments in order to answer this question. For those randomly chosen students, a random initial skill (skill 'A' in Figure 9-2) is selected and the students get to answer questions from the prerequisite skills ('B' and 'C') of the chosen skill if the students get the initial skill correct.

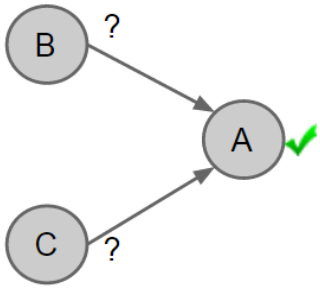


Figure 9-2 Sample navigation of the graph for this study

In order not to overload the chosen students with work, in a given assignment, only one initial skill is used in the study. Additionally, if any of the chosen students gets the prerequisite skill incorrect, they are not assigned remediation assignments as is the case with all the other assignments, and the navigation does not continue to the second level of prerequisite skills (i.e. those of 'B' and 'C' in Figure 9-2) for the chosen initial skill.

It is expected that if a high percentage of the students in the study answer the prerequisite skills of a given skill correctly, this would suggest a strong relationship between the skills, and hence maintain the link in the graph. On the other hand, if the percentage is below a predetermined threshold, it would suggest that that prerequisite link in the graph would either require further scrutiny or must be removed.

9.2.1 Dataset

The dataset includes a prerequisite skill graph developed by a Mathematics domain expert. This graph contains skills from the Common Core Standards [17] spanning grades K-9. The graph, which is the graph used in PLACEments, has a total of 495 prerequisite skill relationships. A portion of the graph is shown in Figure 9-3 below. The green lines in the graph indicate that the prerequisite skills are in a lower grade level, whiles the black arrows between the nodes show prerequisite links between skills of the same grade level as the post-requisite skill. The node names represent the skill codes from the Common Core Standards for Mathematics. Each of these nodes has a complete description and examples of what students need to be taught. See http://www.corestandards.org/Math/ for a complete listing and a detailed description of the standards.

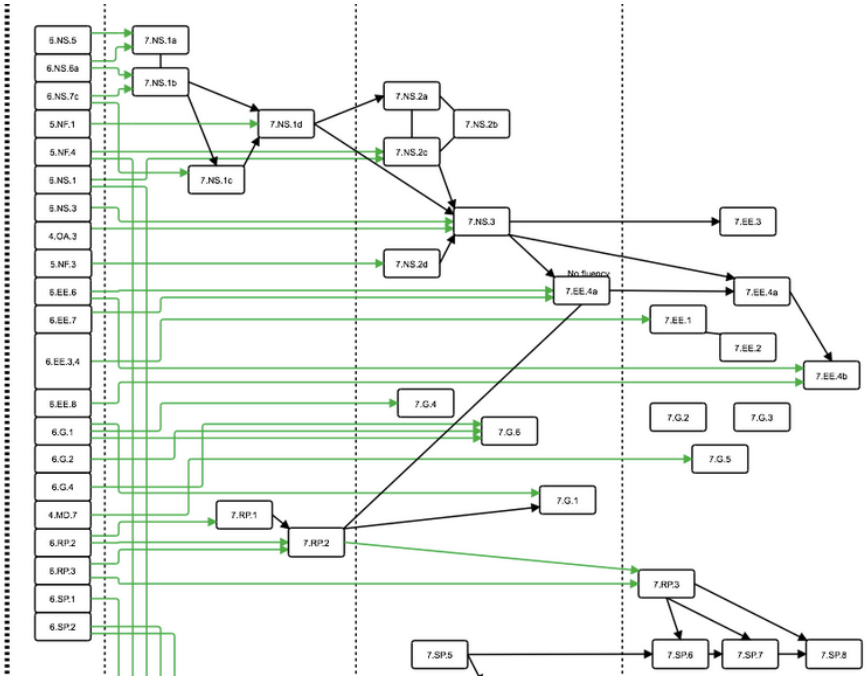


Figure 9-3 A portion of the prerequisite skill graph designed by a math expert and based on standards from the Common Core Mathematics Standards [52]

The dataset additionally includes 1272 problem logs from ASSISTments. Each row in the dataset represents a student's response to a placements test item. That dataset also includes a matrix of item to skill tagging. These logs were from 601 distinct students whose grades ranged between 6 and 12. Each of these students was assigned at least one of the 119 placements assignments used in the study. The data set represented 60 of the 495 prerequisite relationships in the skill graph.

9.3 Results and Analysis

As of the time of reporting this study, data had been collected on 60 of the 495 relationships/prerequisite skill links. Of these 60, 35 had at least 10 responses. (See Table 9-1 for the complete list of 35 links) We limit the number of responses per relationship to 10 in order to achieve some generalization of the results. The graph in Figure 9-4 shows that three (3) of the relationships had link strength of 0, since none of the students had the prerequisite questions correct even though they knew the post-requisite. Two (2) were of the maximum strength (i.e. 1). A larger proportion of the links examined so far has strengths ranging between 0 and 1. As many as 24 of the links have a significantly low link strength as the figure shows.

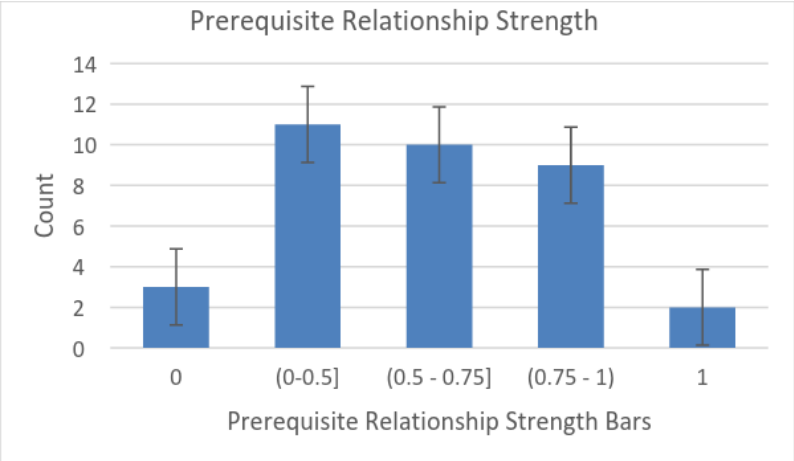


Figure 9-4 Prerequisite Link Strength

As the graph in figure 9-4 indicates, we can make general statements about the relationships. For three of the relationships, none of the students knew the prerequisite skills even though they performed well on the post-requisite skill. Similarly, two of the links can be believed since all the students who knew the post-requisite skills also knew the pre-requisite, suggesting that the link belongs in the graph. There was a larger number of the links for which strengths were inconclusive. Of particular interest are the skill links with strength below 0.5. Those strength values show that a big percentage of students did not know the prerequisite skill even though they all got the post-requisite skills correct. These low numbers suggest that the prerequisite relationship between the skills need to be looked at extensively and may warrant a removal from the skill graph. It may be safe to assume that the skills with a link strength above 0.5 may be valid and the reason for which the strength is not 1 may be because the items used in the test have high slip rates. However, this assertion needs further studies to ascertain.

9.3.1 Effect of Student Knowledge Levels on Link Strength

To help us understand how the link strength is affected by the knowledge level of the students who participated in the study, the results were subdivided into the different knowledge levels. The knowledge level of a student was determined by the student's prior percent correct, i.e. the percent of correctness of a student's previous performance on problems in ASSISTments prior to the study. All students with a percent correct value below 0.5 were assigned to the low knowledge group, while those students with percent correct values between 0.5 and 0.75 were tagged as medium level students. Any student whose prior percent correct was above 0.75 was tagged as a high knowledge student.

Figures 9-5 and 9-6 present a breakdown of the results by knowledge level. Table 9-1 lists the 35-links considered in this study. Of the 35-skill links studied, there were 12 of the skills for which we had data for all three knowledge levels. Twenty-three (23) of these links were examined for only medium and high knowledge students.

The results, in Figure 9-5, show some variations in the link strength when the data is split into different knowledge levels, with the exception of two of the links. The results for links J and K clearly show an agreement in the results across knowledge levels. The results for J suggest a very weak link and hence that link must be removed from the skill graph. For all knowledge groups, link K is strong, suggesting that the link is believable and hence belong in the graph.

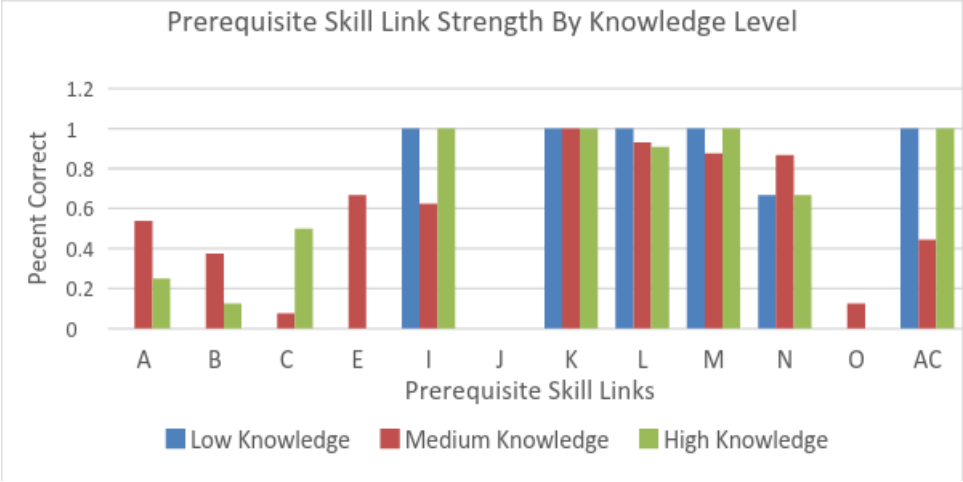


Figure 9-5 Prerequisite Skill Link Strength by knowledge level

Table 9-1 A Subset of the List of Skill Links in the prerequisite skill graph.

Link Code	Skill	Prerequisite Skill
A	Ordering Fractions	Equivalent Fractions
B	Subtracting fractions like denominator	Adding mixed numbers like denominator
C	Comparing Positive Decimals	Read & write decimals
D	Subtraction Mixed Numbers	Addition Mixed Numbers
E	Word problems with fractions as division	Multiplication Fractions
F	Multiplication Fractions	Area of rectangle word problems
G	Line Plot	Real world fraction multiplication
H	Line Plot	Line Plot with fractions
I	Expressing unit rate in words	Finding the Ratio
J	Expressing unit rate in words	Word problems with fractions as division
K	Solve unit rate problems	Expressing unit rate in words
L	Percent of	Expressing unit rate in words
M	Divide multi-digit numbers	Division Whole Numbers
N	Division of Positive Decimals	Multiplication Positive Decimals
O	Comparing integers on number line	Plot on coordinate plane
P	Evaluate exponents	Multiply by Powers of 10 (number of zeros)
Q	Deviations in measures of center & spread	Median
R	Identify constant of proportionality	Unit Conversions with ratios
S	Identify constant of proportionality	Unit rate with fractions
T	Identify constant of proportionality	Solve unit rate problems
U	Identify constant of proportionality	Percent of
V	Identify constant of proportionality	Expressing unit rate in words
W	Identify constant of proportionality	Percent- finding whole
X	Divide Integers	Multiply Integers
AA	Word problems all operations w/ integers	Word problems with fractions as division
AB	Word problems all operations w/ integers	Divide Integers
Y	Word problems all operations w/ integers	Multiply and Divide non-integer rationals
Z	Word problems all operations w/ integers	Multiply Integers
AC	Combining Like Terms	Distributive Property
AD	Equation Solving Two or Fewer Steps	Word problems all operations w/ integers
AE	Scale drawings	Identify constant of proportionality
AF	Operations with scientific notation	Dividing Monomials
AG	Operations with scientific notation	Power of Powers
AH	Operations with scientific notation	Multiplying Monomials
AI	Transversal	Sum of angles

Links I, L, M, N and AC show that the different knowledge levels contributed differently to the strength. While for links I, M and AC, both high and low knowledge students demonstrate that those links are strong, there is a different result for the medium knowledge students. Apart from links I and AC, all the other links in that

group have a link strength above 0.8. The other interesting link is C. Though the results show link C as a weak link, it is a much weaker link for medium knowledge students than for high knowledge students, and worse for low knowledge students. Another set of interesting results was that for links A, B, E and O. In all of these, it would be expected that medium knowledge students will do poorly on a prerequisite skill than high knowledge students, however the results show that this is not the case for those skills. The medium knowledge students performed better than expected. Link O is even more interesting. The data for both low and high knowledge students suggest that the link should be removed from the skill graph. However, this is a much weaker statement to make for medium knowledge students, suggesting that there should be different prerequisite skill graphs for students with different knowledge levels.

In the data set, there were 23 of the links for which we did not have responses from low knowledge students. The medium and high knowledge level students are compared in Figure 9-6 and we can see that each of the two knowledge groups contributed differently to the results of the study. Links AG and AH appear to be non-existent since the results show that none of the students in the two knowledge groups demonstrate proficiency in the prerequisites in that link even though they responded accurately to the post-requisite skill's question. These links appear to be other candidates for removal from the skill graph. The results in Figure 9-6 show a set of interesting variations in the link strengths across the knowledge levels.

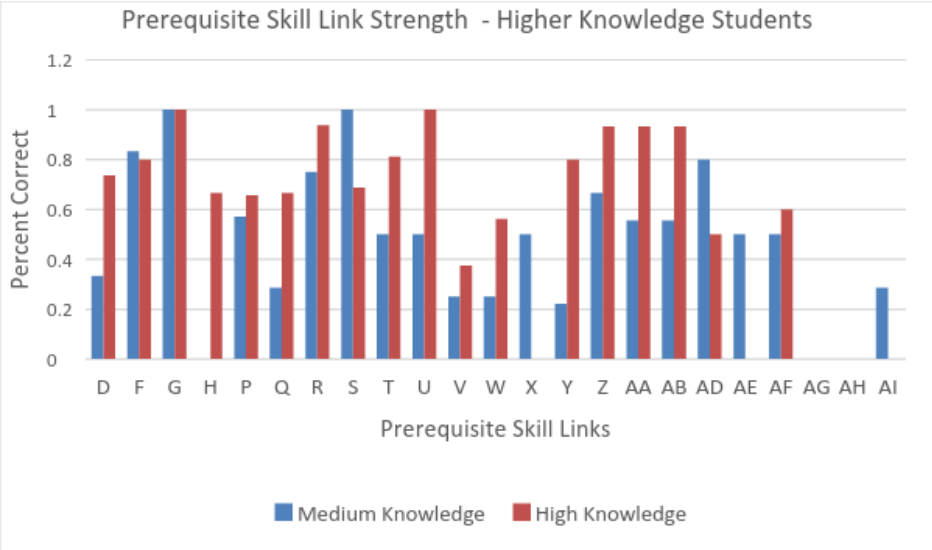


Figure 9-6 Medium and High Knowledge Students' contribution to link strength

Overall, these results suggest a number of the prerequisite skills that need to be assessed in the graph: some may require complete removal from the skill graph and others suggest a different skill graph for different students with different knowledge levels. Additionally, breaking the results down into the different knowledge levels has resulted in one minor finding: Students with different knowledge levels have different representation of knowledge and hence different skill graphs need to be designed for students with different knowledge levels.

9.4 Limitations of the Approach

The approach described in this chapter has a number of limitations. The first limitation relates to the choice of questions for the skills in the skill graph. Since the current implementation of PLACEments uses just one item (or question) to represent a skill in the test, a poorly chosen problem will affect the performance of the search. If a problem chosen to represent a skill in the test has a very high slip or guess rate, the performance of the students (the basis of which is used to determine the strength of links in the graph) will be affected. In other words, if a problem has a high slip rate, PLACEments will assume that students do not know a skill because most of the students will not perform well on that problem, even though there might be a high probability of the students knowing the skill tagged by the problem. An additional limitation that relates to the choice of the questions for skills in the graph is about the number of questions used in the test to assess a student's knowledge of a skill. Since PLACEments currently uses just one question per skill, the choice of problems has to be such that it is a good representation of problems tagged by that skill. In other words, it has to be an almost perfect determinant of a student's knowledge of the skill. One way to deal with this limitation is to use multiple problems with varying guess and slip rates for a given skill.

Another limitation of this approach is that the fact that we start the search by believing the initial set of skills and their ordering. In fact, the approach does not help in determining whether the ordering of the skills in a skill graph is problematic or not. A new ordering of skills cannot be suggested by using this approach. Finally, a large number of student data is needed in order to make reasonable conclusions.

9.5 Contribution

In this chapter, we have proposed an intuitive but novel method for improving prerequisite skill graphs. The freely available adaptive testing system, PLACEments, can be used to collect and analyze student performance data on the items tagged by the skills in a skill graph in order to determine the appropriateness of some of the skill links in a given prerequisite skill graph. Of course, with the limitations mentioned earlier, we think that the educational data mining community can take a good look at this process and augment the search for better fitting models with this new technique.

We have also shown that students of different knowledge levels learn differently and as such there should be a different representation of the skills their learning trajectory in a given domain. Our results suggest that curriculum designers may want to think about the needs of these different knowledge levels in the design of curricula.

9.6 Conclusion and Future Work

Several methods have been proposed and used to improve upon the predictive power of learning maps. These methods include the Learning Factors Analysis, Q-Matrices and the greedy search algorithm proposed and reported upon by the authors in an earlier paper. Many of them have shown promise, especially with the greedy search algorithm showing that an equally predictive model can be found which is different from the initial skill graph but with fewer skill-nodes. [2] However, we have not found any empirical studies that have been conducted with the aim of determining the flaws in a given skill graph/learning map and fixing them. The present study set out to solve this issue. The initial research question was: Can we do a better job at determining the deficiencies of a skill graph with empirical data and improving upon the skill graph?

To answer that question, we built an adaptive testing feature, PLACEments, in ASSISTments and used that to collect data on prerequisite skill graphs. The results of the study showed that deficiencies of a skill graph can be determined. Some relationships can be identified as unnecessary and hence be removed from the system. This finding is true irrespective of the knowledge level of the students involved. Additionally, we found that students with different knowledge levels of a given domain require different skill graphs. This was from the fact

that some of the relationships appeared to be stronger for students of one knowledge level than for those of another.

There were a few limitations of this approach requiring further studies to make the method more robust. It was mentioned that just one item was used to test a students' knowledge of a skill, and so our estimate of the students' knowledge of a skill may be biased by either the difficulty of the item or the guess and slip rates of the item. This requires further studies. One particular future study to perform in this regard is to vary the number and difficulty of the items used to estimate a student's knowledge of the skill.

The results would be made even stronger with more student data. So far, we have looked at only 35 of the over 400 different links in the graph. More data is needed to test different links in the graph. We have therefore proposed an alternative method for improving upon prerequisite skill graphs. This method can be used to augment the results from the other three methods mentioned earlier.

10 Sequencing Content in an Adaptive Testing System: The Role of Choice

The effect of choice on student achievement and engagement has been an extensively researched area of learning analytics. Current research findings suggest a positive relationship between choice and varied outcome measures, but little has been reported to indicate whether these findings hold in the context of Intelligent Tutoring Systems (ITS). In this chapter, we report the results of a randomized controlled experiment in which we investigate the effect of student choice on assignment completion and future achievement in an ITS. The experimental design uses three conditions to observe the effect of choice. In the first condition, students are able to choose the order in which to complete assignments, while in the second condition, students are prescribed an intuitive order in which to complete assignments. Those in the third condition were prescribed a counter-intuitive order in which to complete assignments. Results indicate that allowing students to choose the order in which to work on assignments leads to higher completion rates and better achievement at posttest. A post-hoc analysis also revealed that even considering students with similar completion rates, those given choice had higher posttest scores than those observed in any other condition. These results seem to support the many theories of the positive effect of choice on student achievement.

This chapter is submitted to the following venue:

Seth A. Adjei, Anthony F. Botelho, and Neil T. Heffernan. (**2017**). Sequencing content in an adaptive testing system: the role of choice. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (LAK '17). ACM, New York, NY, USA, 178-182. DOI: https://doi.org/10.1145/3027385.3027412

10.1 Introduction

The concept of mastery learning is based on a philosophy that states that "all students have the ability to learn anything" and that this ability is a function of time. In other words, given a new topic, it is merely a matter of time and practice before one can reach a state of understanding. It has also been suggested that mastery learning is purely teacher-paced, where teachers determine the order in which students must learn specific knowledge components or skills.

An opposing philosophy to mastery learning is known as the personalized system of instruction (PSI), in which students decide on their pace and the amount of content they learn.[13, 75] conducted a study in which mastery learning of content in the Cognitive Tutor was compared to teachers' prescriptions of the order in which to present content.[75] It was found in this work that the system's determination of ordering caused significant improvements in student learning. Their findings suggested that using ITS to prescribe the order in which students were presented a set of knowledge components or skills was a better approach to learning than allowing teachers to determine or prescribe content order. From a different perspective, these results seemed to show that choice, at least at the teacher level, did not cause learning gains.

The effect of choice on various aspects of human life has been studied for many decades. Watanabe & Sturmey performed a meta-analysis of publications in the area of metacognition and the effect of choice on student performance and found that, particularly for students with disability, allowing choice has many benefits. [96] Choice was shown to improve student engagement on tasks as well as propensity of completion. Additionally, it has been shown that intrinsic motivation to carry out general tasks can be improved when students are given choice. [27] Other researchers have observed the positive effect of choice on outcome measures in a number of varied activities. [50, 66, 101] Wang & Stiles showed that students completed more tasks when asked to choose when and how to complete the tasks. [94] This phenomenon is evident in preschoolers [6], high-schoolers [69], and college undergraduates [101]. Across ages, the primary contributing factor causing the increase in performance and rate of completion has been attributed to the motivational effects of choice.

The extent of the effects of choice are sometimes conflicting across different studies. Flowerday and Schraw [38], for example, show that choice had a positive effect on attitude and effort, however the effect on cognitive engagement was minimal or non-existent. While not unanimous across all domains and studies, there is compelling argument to pursue the study of choice for its potential benefits in learning. Understanding how the positive aspects of choice can best be implemented to improve students' learning experiences is a topic still in need of research.

Despite the many benefits that seem to be derived from choice, ITSs rarely offer features that allow students to make choices regarding what they learn, and when and how to remediate content that they may be lacking. Ostrow & Heffernan conducted a randomized controlled experiment in which they investigated allowing students to choose the type of feedback received while working on an assignment and its effect on assignment completion and future performance. [62] They compared students who were given the choice to decide on the type of feedback received with those who randomly received a particular type of feedback. They found that students given choice had significantly better achievement than those in the control group, lending credence to the notion that choice has a positive impact on student performance within an ITS.

In this study, our goal is to investigate the effect of choice on student assignment completion and learning gains when given the opportunity to choose the order in which to complete assignment tasks. We report on a randomized controlled experiment in which students were placed into three conditions. In one condition, students were asked to choose the order in which to complete the assignments, whereas students in the other two conditions followed different prescribed content orders. We also report a post-hoc analysis of the study in which we find that, for students with similar assignment completion rates, those in the choice group performed better at posttest than those in either prescribed condition.

10.1.1 Research Question

The following research question is addressed in the present study:

- *Does allowing students to choose the order in which to remediate skills improve adherence in the form of assignment completion rates, and/or Math achievement?*

In other words, does choice matter? What is the relationship between student choice and mastery learning?

10.2 **Methods**

This section describes the methods employed in answering the research questions stated above. We ran the randomized controlled experiment in PLACEments, an adaptive testing system. This system is described briefly in this section. We then present the experimental design, the participants used in the experiment, and the outcome measures of interest.

10.2.1 An introduction to PLACEments

PLACEments is a computer-aided adaptive testing feature of ASSISTments, an online learning platform powered by Worcester Polytechnic Institute. [42] PLACEments uses a prerequisite skill graph that underlies the system, created based on the Massachusetts Common Core State Standards for Mathematics. [17] All PLACEments tests are teacher-driven, in that teachers choose what and when to assign. These tests are composed of an initial set of skills selected by the teacher, and once assigned, students are tested on questions related to the initial skills. If a student performs poorly on any of the initial skills, the system traverses the skill graph to select questions from the immediate prerequisite skills of the incorrect items. These items are then included on the test, and the graph traversal for item selection continues until the system determines that there are no further prerequisite skills to be shown or the traversal reaches a predefined end point in the graph; the predefined end-point is set at test creation time. In this manner, the system can isolate and map the depth of gaps in students' knowledge, while providing opportunity for remediation.

10.2.1.1 Progressing through the Test

For the sake of simplicity, we use the hypothetical graph shown in Figure 10-1 to explain how the test proceeds. The nodes in the graph represent skills or knowledge components. The arrows between skills represent the order in which students need to learn these skills/concepts in order to succeed in the subsequent skills. They therefore show the prerequisite relationships between skills (thus, skill 'D' is one of the prerequisites of skill 'A'). The correctness indicators attached to each node in the graph are a representation of a given student's performance during the test. In this configuration, the student is assigned 'A', 'B', and 'C' as initial skills in the test. The system presents the student with questions from these skills, and since the student performs poorly on skill A (as shown in the graph), the student is further tested on skills D, E, and then subsequently H since the student did not demonstrate mastery of skills E and H respectively.

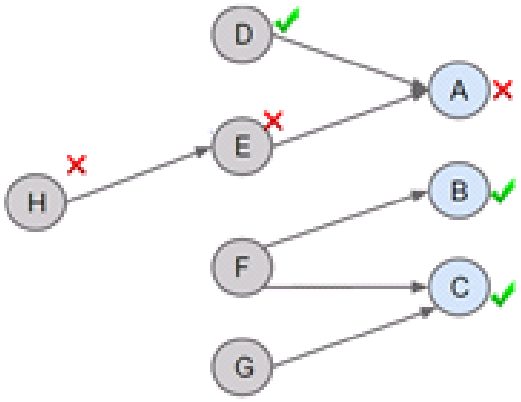


Figure 10-1 A sample skill graph and a sample student's response configuration

Generally, the tests are meant to identify students' lack of specific skill knowledge and to find which prerequisite skills to blame for that missing knowledge.

10.2.1.2 Remediation Assignment Creation and Release

Once the knowledge gaps are determined from the test, PLACEments attempts to help students close that gap. Once the test is completed, students are assigned remedial practice questions on the skills in which they performed poorly. The release of these assignments is staggered and is based on the underlying prerequisite skill graph that PLACEments depends upon, and the number of these "remediation" assignments given is dependent on the student's performance during the test. The remediation assignments of the lowest grade level skills on which the students performed poorly are released first and, once completed, subsequent post-requisite skill remediation assignments follow. In the configuration depicted in Figure 10-1, the remediation assignment for skill H is released and completed before the assignment for E is released. The assignment for skill A will be held back until the student completes skill E. All remediation assignments are mastery-based assignments referred to as "skill builders," in which students are given similar skill-based items until a predefined threshold of understanding is reached; this threshold is usually met by answering three consecutive items correctly.

10.2.2 Experiment Design

We ran a randomized controlled trial in PLACEments in which we experimented with the order in which remediation assignments were released. Figure 10-2 illustrates the experimental design for this study.

As shown in Figure 10-2, each participant is given a predefined PLACEments test which has various initial skills. These assignments are teacher-assigned and may have varying degrees of difficulty. After the tests, students are randomly placed in one of three conditions. In the first condition, "Prerequisite to Post-requisite," participants are assigned remediation skill builder assignments beginning with the skills of the lowest grade level and the graph is traversed in the pre-to-post direction. Participants are required to complete all released remediation assignments for a given test before the subsequent post-requisite skill related assignments are released. This condition typifies the current graph traversal direction for remediation assignments that are released in PLACEments (See section 10.2.1.2 for more details).

The "Post-requisite to Prerequisite" condition has a similar behavior as the "Prerequisite to Post-requisite" condition with the exception that the graph is traversed in reverse, from the post-requisite to the prerequisite skills, which is counter intuitive to most teachers. In the third condition, graph traversal is not considered. For all participants in this third condition, the release of remediation assignments is not staggered, nor is it based on the prerequisite skill graph. Instead, all remediation assignments are released to the students at once and they get to choose the order in which to complete their assignments. A month after the initial test, students had the opportunity to retake their initial PLACEments test as a posttest to gauge the amount of learning that had occurred from the remediation assignments and, ultimately, the effect of condition.

We also performed a post-hoc analysis of the data collected from this experiment. In this post-hoc analysis, we investigated the effect of other PLACEments test features and the condition assignment on student's performance gain over the study period.

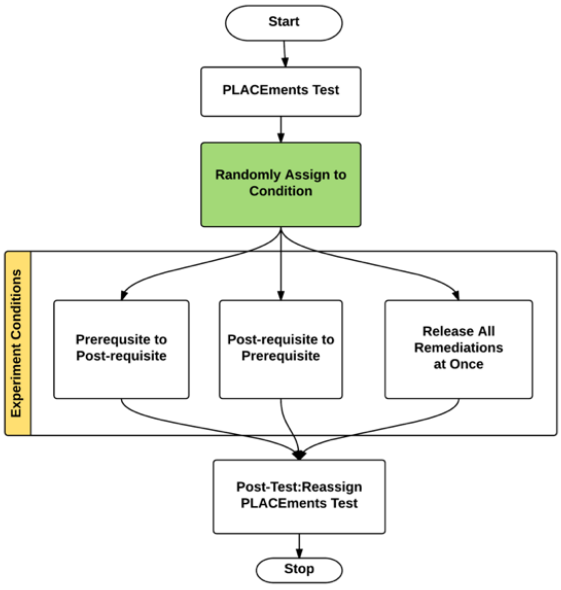


Figure 10-2 Experimental Design

10.2.3 Participants

For this experiment, there were 410 student participants, each of whom was assigned the initial PLACEments test as well as the reassignment that served as the outcome measure. All students were 7th and 8th grade users of ASSISTments. The participants had varying levels of math competence and were randomly assigned to one of the three previously described conditions in the study. The "Prerequisite-to-Post-requisite," "Post-requisite-to-Prerequisite," and "Release All" conditions had 129, 145, 136 students respectively. Random assignment to condition was performed after the initial PLACEments test was completed. The results of the tests in no way impacted random assignment.

10.2.4 Outcome measures

To determine the effectiveness of choice, the following outcome measures were used: *remediation completion rate*, *performance on posttest,* and the *learning gain* from the initial to the reassigned PLACEments tests (i.e., from pre- to posttest).

The completion rate, in this context, is the ratio of remediation assignments completed to the number of remediation assignments assigned. This outcome measure was intended to help determine whether the order in which remediation assignments were released had an impact on students' assignment completion rates.

Additionally, we use students' performance on the second PLACEments test as a second outcome measure (i.e., posttest). We also considered the gain in PLACEments test performance. This gain was the mathematical difference between the initial test performance (expressed as percent of items answered correctly) and that of the second PLACEments test.

10.3 **Results**

In this section, we present an initial intent-to-treat analysis of all the participants in the experiment and further describe an analysis of students who participated in the post-test. We then proceed to answer the proposed research questions using data from students who were actually treated. The dataset for this experiment can be found at http://tiny.cc/palsrct5data.

10.3.1 Effect of Choice on Remediation Assignment Completion Rate

Though all 410 students in the study were expected to complete the posttest, we found that a high percentage of students did not have the opportunity to do so. In some cases, teachers prevented entire classes from completing the posttest, while in other cases, the school year ended before students had the opportunity to take the posttest. In view of this, only one of our research questions can be answered using the entire population of the study.

In regards to the impact of choice on assignment completion**,** Table 10-1 shows the remediation completion rate for each of the conditions in the study. There was no significant difference in remediation completion rates between conditions (p-value > 0.05). Though students in the counter intuitive condition (i.e. post-to-pre condition) seemed to have a slight edge over students in other conditions, the difference was not significant. The observed difference may be due to the fact that the post-to-pre condition encouraged students to complete more assignments because they presumably navigated from difficult assignments to easier assignments. Generally, there was a low average remediation assignment completion rate of 0.38 across the entire population.

Table 10-1 Remediation Completion Rates by Condition

Condition	Participants	Mean Completion Rate
Pre-to-post	129	0.38
Post-to-pre	145	0.42
Student Choice	136	0.35

10.3.2 Effect of Choice on Post-test Performance

Of the 410 initial students, 70 students completed the post-test. As Table 10-2 shows, the students were randomly and almost equally distributed across the different conditions. This section describes the effect of choice on performance of these students on the post-test.

Table 10-2 Completion Rates and Learning Gains

Condition	Number of Participants	Average Completion Rate*	Learning Gain*
Pre-to-post	22	0.457	0.038
Post-to-pre	26	0.476	0.120
Release All/ Student Choice	22	0.512	0.310
Total	70		

* Significance with p-value <0.05

Among others, Table 10-2 shows that students in the pre-post condition who completed the posttest also completed far more remediation assignments than those in the other two groups. Additionally, students in the choice condition were not completing as many assignments as those in either prescriptive condition. These results suggest that choice in this setting did not necessarily increase assignment completion rates for these students, as described in section 9.3.1 above. However, Table 10-2 suggests that this same group of students performed better on the posttest than students in the two prescriptive conditions. Their gain in achievement from the pre-test to the post-test was more than twice the gain for the counter intuitive group and 10 times that of the intuitive group. This seems to suggest that students in this condition may have recognized the skills they performed poorly on and were therefore able to make intelligent choices regarding which skills required remediation.

We also performed a one-way ANOVA and the results show that there was a significant difference in math achievement for students who had the chance to complete the experiment (p-value < 0.05).

10.4 **Post-hoc Analysis of Results**

Across our population, 32 students had an assignment completion rate of 100%. (see Table 10-3) We analyzed these 32 participants and found that, first of all, they were equally distributed among the conditions. Secondly, students in the choice condition achieved huge learning gains over those from the other prescriptive conditions. This result seems to suggest that even among students who are consistent in completing their assignments, prescribing the order in which to complete assignments is not ultimately helpful to learning. When there are multiple tasks to be performed by students, it is best to allow them to choose the order in which to work on the assignments, as suggested by our results here. Allowing students to choose the order in which they work on assignments appears to provide better gains than when the systems make the choice for them, especially for students who have high assignment completion rates.

Table 10-3 Learning Gains among students with comparable assignment completion rates

Condition	Number of Participants	Learning Gain*
Pre-to-post	11	0.056
Post-to-pre	11	0.086
Student Choice	10	0.333
Total	32	

* Significance with p-value <0.05

10.5 **Discussions**

Student choice has been found to be helpful for encouraging learners to perform well on certain outcome measures of interest. Research has shown that giving students opportunities to make choices regarding the pace and sequence of math content has many positive effects on students. These findings informed our quest to determine whether the phenomenon would hold true in the context of PLACEments, the adaptive testing system that leverages the ASSISTments learning platform.

In the current study, we set out to determine whether the touted benefits of student choice could be replicated in our testing system, and if so, to what degree it mattered. Contrary to the established notion that choice improves assignment completion, the present study showed that assignment completion rates were not significantly different among conditions. These findings reveal that though there were differences in student completion rates, these differences were not significant and their magnitudes were minimal at best. We think this may be the result of several factors, the most prominent of which is the possibility that the lengths of the PLACEments tests in these classes were too short.

However, of the students who completed the posttest, we found that differences in assignment completion were significant. We also found that among those students who completed the experiment, there were significant differences in learning gains. Post-hoc analysis of the results seemed to suggest that choice was very important amongst students with comparable assignment completing behavior. This is an impactful finding, as it suggests that choice increases performance. Of note here, the observed performance boost could not be attributed to students completing more assignments than those in the other groups; the assignment completion rates were not significantly different, and yet the difference in performance remains.

The contributions of this chapter support that in every learning analytics study that tries to model students learning and behavior, the effect of choice cannot be ignored. Additionally, designers of ITSs must look for ways to incorporate opportunities for students with comparable abilities and assignment completion rates to make choices in the order in which they complete assignments while using the system. This consideration will contribute to an improved learning experience for students.

10.6 **Future Work**

The study we report in this chapter has one clear limitation in that there was a considerably high dropout rate among all experimental conditions. We presume this high attrition may have been caused by a number of possible factors which require further scrutiny. We think that this may be an artifact of the PLACEments system and the size and difficulty of the assignments used in the study. Further investigations into the causes of this high dropout

rate are necessary to help rectify the issue in future analyses, and to boost teacher and student fidelity of the PLACEments system.

This chapter reports the results of investigating the effect of choice on the release and completion of remediation assignments. Another feature of the system in which we can implement choice is in the test itself. Additional experiments are being planned to determine how choice can be incorporated in this aspect. An illustrative example of this involves providing choice in completing the initial skills for the test.

We intend to run additional experiments to replicate these findings and improve upon the current results. If the results hold in replication trials, we will modify the PLACEments system to allow students to choose the order in which to complete their remediation assignments as it is shown here to significantly benefit student learning.

11 Refining Prerequisite Skill Links using Randomized Controlled Experiments in PLACEments

11.1 Introduction:

According to the National Assessment of Education Progress (NAEP), the rate of growth of the performance of 4th and 8th grade students in math and reading assessments in the United States has stalled over the past few years. [60] This finding was consistent across almost all demographics and is possibly indicative of the fact that there is room for improvement in student achievement or is evidence of a ceiling effect (i.e. the curriculum is just as good as it is going to get). Identifying the major causes of such a stalling is very important to ensure that students get the most out of the education they receive at school. Better yet, creating innovative solutions to address discovered causes will then help to ensure that instructional practices grow to benefit student learning. The importance of taking these steps cannot be overemphasized, since research has shown that students' performance in math, among other factors, is highly correlated to their performance on the job in future [83].

As was stated earlier in in section 2.2, learning is defined as "the activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something" (Webster's Dictionary). This definition shows that learning requires a number of strategies. Many of these strategies have been categorized into two areas: primary and support. [74] The primary activities include identification, comprehension, retrieval, and utilization strategies, while the support techniques relate to the activities that support the primary activities. In the words of Dansereau [74], these activities "allow the primary strategies to flow efficiently and affectingly." They include techniques that ensure that students learn in the right environment, techniques for monitoring the students' progress while correcting primary strategies and, any activity that helps to deal with loss of concentration and develop positive attitudes towards learning.

In spite of the fact that these learning strategies are applicable to learning math and are meant to make math learning interesting and easy, many students continue to struggle with the subject. Several factors have been attributed to this observance. Among others, math anxiety [10], socio-economic factors [55], parental over-expectation [59], prior low math achievement [56], lack of prerequisite knowledge [33] and the poor ordering of instructions have been identified to be major causes of such poor performance in math. This is reflective of the

multi-facetted nature of the problem. Many of these factors are beyond the control of educators. However, the factor to which researchers have given little consideration is the problem of the deficit of appropriate prerequisite skills from a perspective of mitigation and to ensure that the right ordering of skills is followed during instruction. These orderings of skills for a given subject domain (e.g. Math) are designed by subject domain experts. However, over time, performance data has been collected from students who have been taught by teachers who follow these expert-designed learning trajectories. The availability of these sets of data opens up the opportunity for data mining techniques to be used to infer prerequisite skills relationships.

In this dissertation, we tackle a factor that has been little studied: students' lack of prerequisite skills. The earlier chapters of this dissertation have shown the many different data mining approaches that we and other researchers have employed to improve learning trajectories (or prerequisite skill relationships between skills) using learner performance data. The findings presented herein, and those of other researchers who have addressed this issue demonstrate that some progress has been made towards achieving the goal of assisting domain experts to design curriculum that results in the best learning gains for students. As the previous chapters have reported, some of these approaches include Desmarais' Partially ordered knowledge structures [28, 31], Learning factors analysis [19, 63] and deep learning-based methods [67, 100]. While these methods seem to identify and refine prerequisite skill relationships amongst skills, these findings are purely data-mining based and are therefore largely correlational. Very minimal causal claims can be made based on these findings without the use of randomized experimentation. While the earlier approaches appear to have some success, there is a lot more benefit in using randomized controlled trials (RCTs) to infer prerequisite skill structures. [88]

This chapter presents an overview of the PLACEments infrastructure in some more detail, the types of experiments that can be executed within the system, and the results from a number of experiments for which data has been collected.

11.2 **PLACEments**

Chapters 8 and 9 gave brief overviews of PLACEments, an adaptive testing and remediation system. The overall objective of this system is to identify gaps that exist in students' knowledge, and to provide a means by which these gaps can be closed. The system traverses a prerequisite skill structure as it determines gaps in student

knowledge, identifiable as seemingly non-acquired knowledge components, and tries to remedy such knowledge gaps. As supplement to the descriptions given of the system in chapters 8 and 9, a more-directed overview of the system and its features is presented in this section.

11.2.1 Assignment Creation

Typically, the system allows teachers to assign tests, in which they specify knowledge components (or skills) that they want to test their students on. These skills are referred to as initial skills. As students go through these tests, they are assigned test items from the prerequisites of the initial skills they are not able to demonstrate knowledge of. The idea is to identify the prerequisite skills to blame if students perform poorly on a given skill. There are two ways by which the test is completed by the students. In the first case, the test terminates if there are no further prerequisite skills to ask students of, and the student has been tested on all the initial skills. In the second situation, teachers who assign these tests indicate the minimum grade level beyond which students are not to be tested. When a student reaches that point in the prerequisite skill graph traversal, the test terminates.

11.2.2 Filling the Knowledge Gaps

After the test is completed, the set of skills that students could not demonstrate knowledge of is then used to create skill-builder assignments. These assignments are meant to help the students build the knowledge components (skills) that they could not demonstrate mastery of. They include a set of questions created using the knowledge component. These questions are asked of the students and if they fail to answer any of the questions correct, they are given tutoring in the form of hints texts and video instructions. Students need to answer a predefined number of questions correct in a row to achieve mastery. The remediation assignments are then released to the students in a staggered fashion, where by default the assignments of easier skills are released and completed before those of the more difficult skills.

11.3 **Experiment Types Available in PLACEments**

The assessment and remediation features of PLACEments allow for multiple experiments to be run to infer skill topologies or at least refine sections of existing prerequisite topologies. Since the system uses an underlying prerequisite skill structure, it presents an enormous opportunity for randomized controlled experiments to be

designed and implemented to help determine the optimal representation of the prerequisite skill structures that causes more learning. There are many different types of experiments that are possible. This section describes some of the different experiment types that are available and the different types of statements to be made about the portions of the underlying prerequisite skill graph experimented with. Figure 11-1 shows a sample prerequisite skill graph that we will use to explain the different types of experiments described in this section. The graph depicts the relationships among a subset of 4th and 5th grade math content from the common core state standards. [17] A detailed description of the experiment types proposed in PLACEments can be found in Table 11-1. Each of the experiments described in this section is analyzed using analysis of variance (ANOVA) on the individual outcome measures.

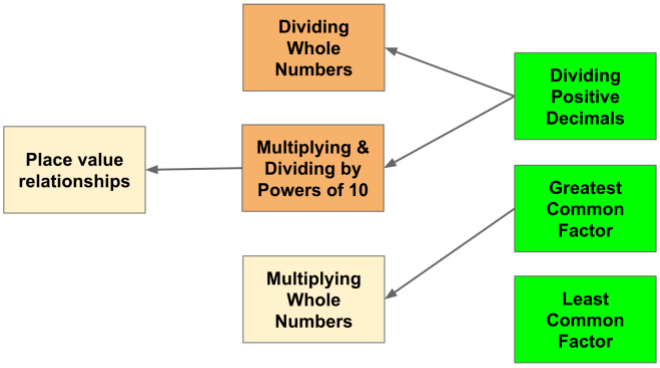


Figure 11-1 Sample Prerequisite Skill Graph.

The arrows emanate from the post-requisite skills and point to the prerequisite skills

11.3.1 Verify Existence of Links

This type of experiments is meant to verify the existence of a prerequisite skill link between two skills. To illustrate, consider the graph in Figure 11-1. According to this graph, "Diving Positive Decimals" has two pre-requisite skills: "Diving Whole Numbers" and "Multiplying and Dividing by Powers of 10." The relationship between one of the prerequisites and its post-requisite can be investigated using the types of experiments described in this section.

11.3.1.1 "Drop Prerequisite Skill" Study

The "drop prerequisite skill" experiment is one that is run in the remediation's feature of PLACEments. In this experiment, we hypothesize that *if there is a prerequisite skill relationship between two skills, then mastery of the prerequisite is necessary for mastery of the post-requisite skill*, and that *mastery of the prerequisite skill should cause faster learning in the post-requisite skill.* In this study, we answer the following research question: "Does practice of prerequisite skills impact students' performance on a post-requisite skill?" The outcome measures of this experiment are the speed with which they complete the assignment for the post-requisite skill and, the amount of assistance and time needed to complete the post-requisite assignment. The completion speed will be measured by the number of problems it takes the student to reach mastery, whereas the amount of assistance sort is operationalized as the number of hints and attempts made when answering any of the questions in the assignment. Participants of this study will be those students who have performed poorly in the test on a pair of skills that the underlying skill graph indicates are related. This is because, these are the participants who have not demonstrated mastery of the skills in the study and who will need additional instruction to demonstrate mastery. Using figure 11-1 as an example, students who perform poorly on both "Diving Whole Numbers" and "Multiplying and Dividing by Powers of 10" will be used. To verify the existence of a prerequisite relationship between two skills, students will be randomly assigned to two conditions. In the treatment condition, students are assigned remediation assignments for the prerequisite skill. They are then asked to complete the remediation assignment of the prerequisite skill before completing the remediation assignment for the post-requisite skill. In the control condition, students are assigned the remediation task for only the post-requisite skill. Participants in this condition will have their remediation assignments for the prerequisite skill marked as completed. Figure 11-2 summarizes the design of this experiment type.

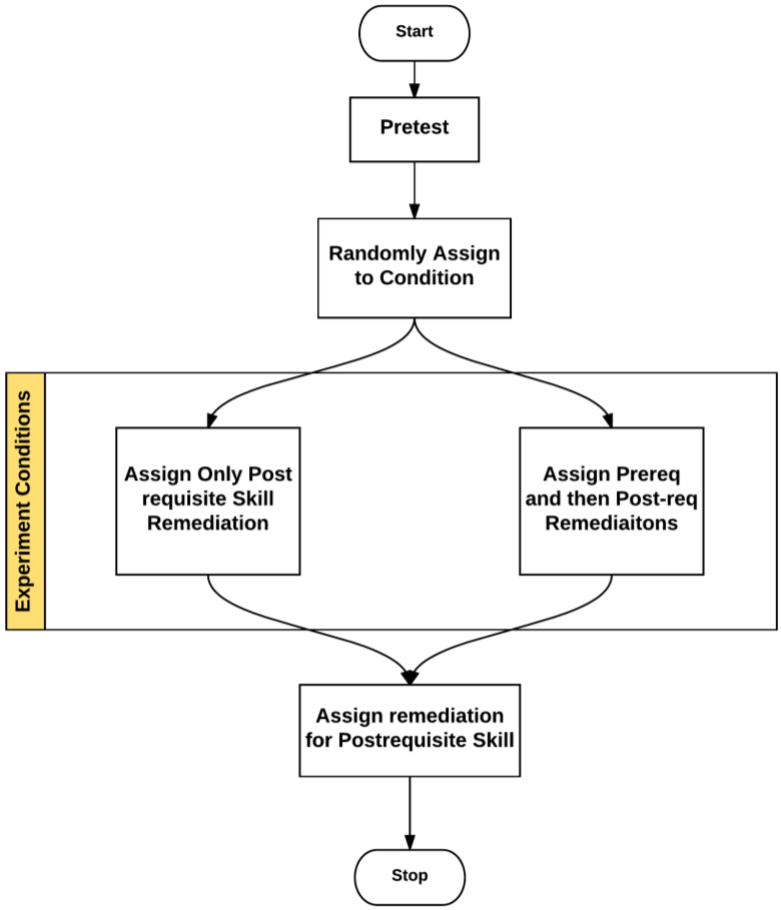


Figure 11-2 Drop Prerequisite Skill Experiment Design

To reduce attrition bias due to an unfair amount of work to be done by the treatment group, we reduce the amount of work by randomly selecting another link for which they need remediation and mark the prerequisite skill of that link as completed. This implies that participants in this experiment must have at least 2 unrelated links for which they need remediation. The following algorithm explains the procedure used for assigning participants to condition. The procedure named "assignToCondition" accepts the list of links (*LP)*, the student performed poorly in during the PLACEments test and the link on which the experiment is being executed (*expLP*). It then checks whether there is at least a link unrelated to *expLP* that can be used in the study. If there is no unrelated skill link which can be used, the participant exists the study. In the event that there is more than an unrelated skill links, one of them is selected at random. If the participant is randomly assigned to the treatment group, the student is assigned practice related to the prerequisite skill of the link expLP and the assignment of the prerequisite skill of the other unrelated skill link is marked as completed. Participants in the control group are not assigned the

assignment of the prerequisite skill in expLP but are assigned that of the related skill randomly selected from procedure *selectRandomUnrelatedLink*. Doing so ensures that there is a fair distribution of work, and attrition bias is removed. See Algorithms 1 and 2 for the complete pseudocode.

Algorithm 1: Assign participants to condition

```
Input: expLP the prerequisite skill link for the current experiment
       LP the list of skill links for which the student needs remediation.
Procedure assignToCondition (expLP, LP)
    if LP.length < 2 then
        exitStudy;
    end
    controlLP ← selectRandomUnrelatedLink(LP, expLP);
    if controlLP.length == 0 then
        exitStudy;
    end
    randCondition ← random number {0, 1}
    if randCondition == 0 then
        //Treatment
        mark assignments related to prerequisite skill of expLP as
         completed;
    else
        // Control
        mark assignments related to prerequisite skill of controlLP as
         completed;
    end
```

Algorithm 2: Select a random unrelated link

```
Input: expLP the prerequisite skill link for the current experiment
       LP the list of skill links for which the student needs remediation.
Result: randomSkillLink is the selected link unrelated to expLP
Function selectRandomUnrelatedLink (expLP, LP)
    candidateLinks ← LP - expLP;
    candidateLinks ← removeRelatedLinks(candidatePairs, expLP)
    if candidateLinks.length > 0 then
        return selectRandomLink(candidateLinks);
    end
    return []
```

In this type of experiment, students' performance in the post-requisite skill's remediation will be used to compare the two conditions. The differences in condition (as measured by completion of the post requisite assignment, mastery speed and percent correct on the initial items and time spent on task) will be used to answer

the research question for this study. An analysis of variance (ANOVA) for each of these outcome measures is performed and used to compare the two conditions.

11.3.1.2 Assign Prerequisites/Post-requisite Skill Irrespective of Performance

This experiment is technically not a randomized controlled experiment. It can be described as a sampling exercise. The sampling occurs during the PLACEments tests. For a subset of the skills, participants who perform well on the post-prerequisite skills are asked questions on the prerequisite skill. Note that this is contrary to the original design of the navigation of the skill graph. Normally students are not asked questions from the prerequisite skill of a skill for which they have demonstrated learning. Hence this is counter intuitive. We hypothesize that if a prerequisite skill relationship exists between two skills, then mastery of the post-requisite skill is necessary and sufficient for mastery of the prerequisite skills of that skill in question.

For another subset of skills and students, we perform a similar sampling exercise in which instead of asking questions from the prerequisite skills the chosen skill, we present questions from the post-requisite skills. For this sampling exercise, we are trying to determine how well knowledge of a prerequisite skill translates to unknown post-requisite skills.

11.3.1.3 "Change Prerequisite Direction" Study

In these experiments, we intend to verify the direction of prerequisite skill links. We answer the question: *Could the direction of a given prerequisite relationship between two skills be incorrect, though the skills have been determined to have a relationship between them*? In other words, will students learn subsequent skills better if the order of a pair of prerequisite skills in a link is reversed?

This experiment type has two conditions per prerequisite skill link. In the first condition, participants are assigned remediation assignments in the order prescribed by the underlying prerequisite skill graph. Participants in the second condition will be assigned remediation assignments in the reverse order of the link. We will then compare students' performance and speed of completion of second assignment in the link order. The results of the comparison will then be used to answer the research question. If one order causes more learning in the second assignment than the other, then we can determine that the order that causes more learning is the correct order.

To illustrate, the prerequisite skill structure graph in Figure 11-1 shows that "Least Common Factor" is a post-requisite of "Multiplying whole numbers". If this link is chosen for this experiment type, participants in the control condition will be assigned remediation in the order specified by the graph. In other words, participants in the control condition will be assigned "Multiplying Whole Numbers" and then "Least Common Factor," whereas students in the treatment condition will be assigned the remediation in the reverse order. Figure 11-3 depicts the simple design for this experiment.

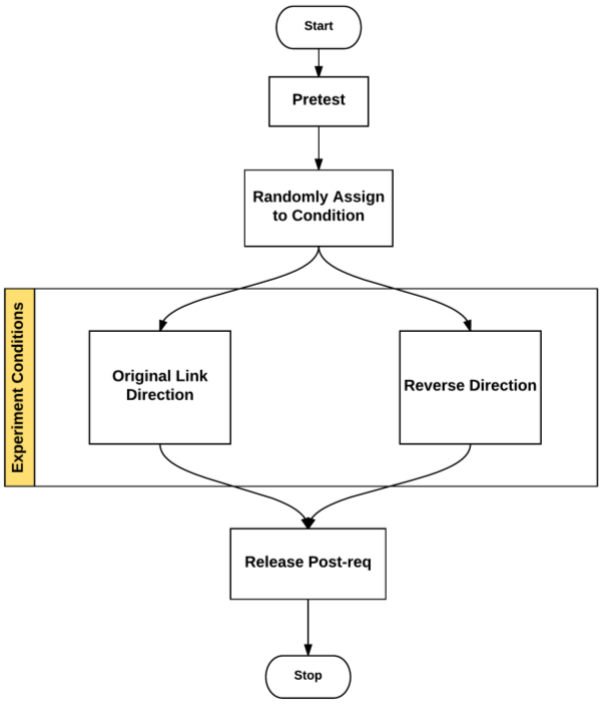


Figure 11-3 "Change Prerequisite Direction" Study design

11.3.2 Find New Links

The PLACEments system allows for another kind of experiments that help in refining prerequisite skill structures. In this set of experiment types, new prerequisite skill links are investigated using randomized controlled trials. In this section, we describe one such experiment type, Order Prerequisite skills experiment.

11.3.2.1 Order Prerequisites

In this experiment, we want to determine whether any pair of prerequisites of a given skill has a prerequisite relationship between them. Specifically, we hypothesize that the order in which students learn two prerequisites of a given skill impacts students' performance on the post-requisite skill, if there exists a prerequisite relationship

between the two prerequisites. In this case, participants will be assigned to two conditions, in which the students get assigned the prerequisites in one of two orders. Using Figure 11-1 as an example, suppose skill "Dividing Positive Decimals" is chosen as a link to experiment with. Students in condition 1 will be assigned remediation for "Divide Whole Numbers" and then the remediation for "Multiplying and Dividing by powers of 10". Participants in condition 2 will be assigned remediation for "Multiplying and Dividing by powers of 10" after which remediation for "Divide Whole Numbers" will be released. The study design is presented in the Figure 11-4 below. Participants will then be assigned the post-requisite skill. Their performance on the post-requisite skills will then be used as post-test for this experiment.

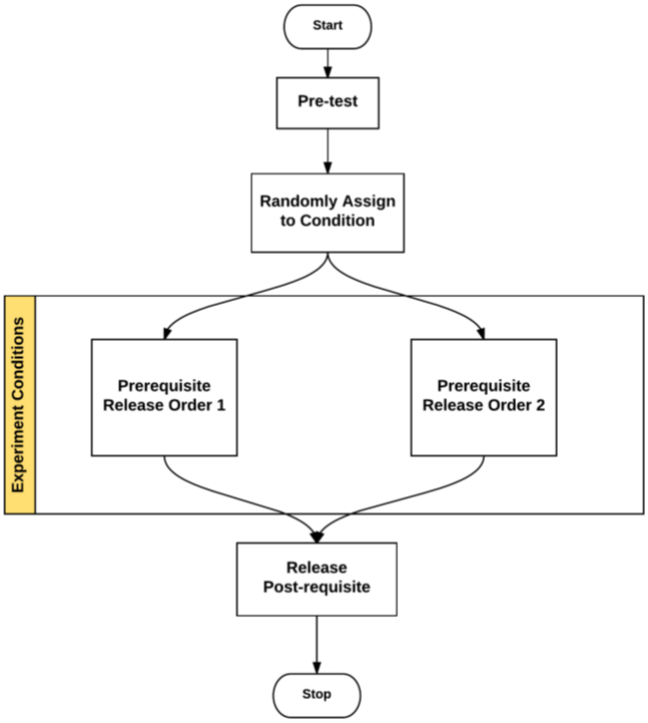


Figure 11-4 Order Prerequisites Study Design

Table 11-1 A summary of Experiments Types available in PLACEments

Experiment Type	Study Name	Research Question	Hypotheses	Outcome Measures	PLACEments Feature
Verify Existence of Links	Drop Prerequisite Skill	Does students' practice of prerequisite skills of a skill impact their learning speed and performance of the post-requisite skill?	*If there is a prerequisite skill relationship between two skills,* • *mastery of the prerequisite is necessary for mastery of the post-requisite skill,* • *mastery of the prerequisite skill should cause faster learning in the post-requisite skill* • *Mastery on the prerequisite skill will improve completion rates of the post-requisite skill*	• Completion of Post-requisite skill remediation • Mastery Speed of post-requisite skill (i.e. the number of opportunities it takes the participant to demonstrate mastery of the post-requisite skill) • Time spent completing the second skill.	Remediation
	Change Prerequisite Direction	Does reversing the order of a prerequisite skill link help participants learn subsequent skills better?	• *If there exists a prerequisite skill link between a pair of skills, reversing the order of the relationship will reduce the speed with which students complete the post-requisite*	• Mastery Speed on the second skill in the link.	Remediation
	Assign Prerequisites	Does mastery of Post-requisite skill imply mastery of prerequisite skill?	• *Mastery of Post-requisite skill implies mastery of prerequisite skill*	• Performance on Prerequisite skill • Time spent on prerequisite skill items (as measured in milliseconds)	Assessments
Finding New Links	Order Prerequisite Skill Pair	For a skill that has two prerequisites, is there a prerequisite relationship between the pair of prerequisites?	• *If there is a prerequisite skill relationship between two prerequisites of a post-requisite, then the order in which the prerequisites are practiced impacts performance on the post-requisite skill.*	• Performance on Post-requisite skills, as measured by mastery speed • Time spent on prerequisite skill items	Remediation

11.4 **Refining the Underlying Graph of PLACEments**

The PLACEments system uses an underlying prerequisite skill graph for middle school mathematics. This graph was developed by domain experts based on the common core standards for mathematics [17]. The complete version is made up of 130 skills involved in 165 prerequisite skill relationships. Some of the skills had no prerequisites, whiles others had up to 5 prerequisites. Figure 11-5 shows a small section of the graph. The rectangular boxes represent skills and the arrows between them represent the prerequisite skill relationship. The arrow points from the prerequisite skill to the post-requisite skill. We therefore run the four experiment categories described in section 11.3 on this prerequisite skill graph structure.

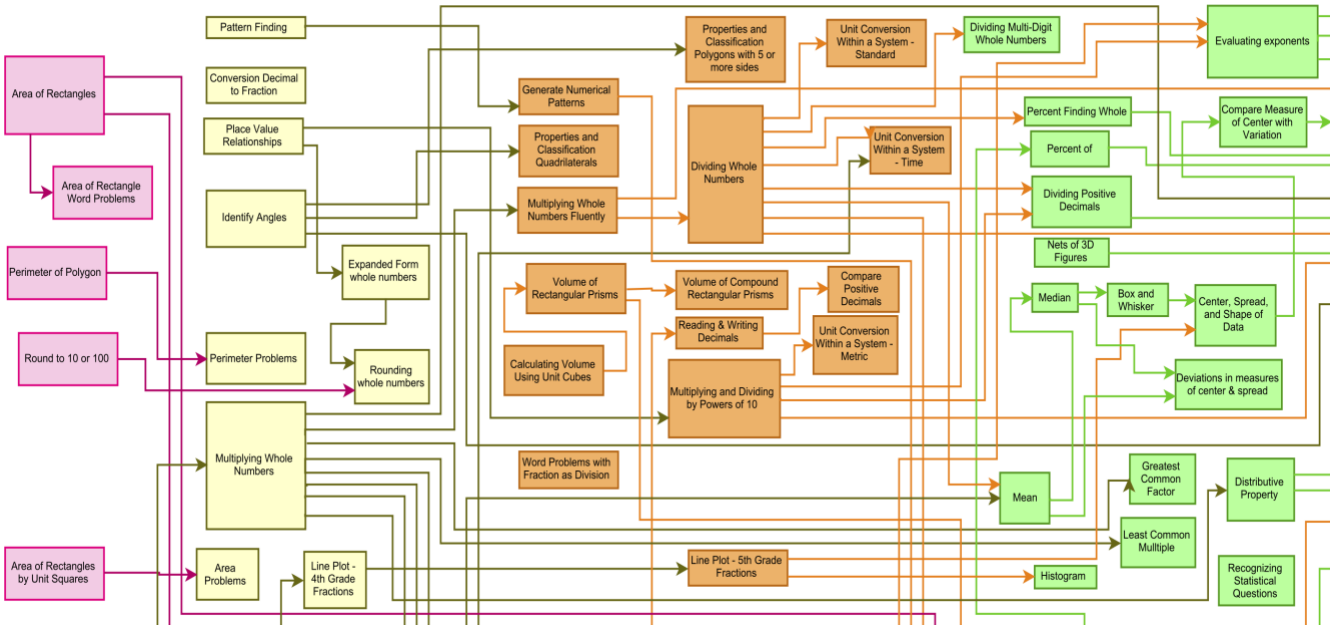


Figure 11-5 A Section of the PLACEments Skill Graph for Middle School Mathematics

11.5 **Results**

Over a period of one academic year, data collected for each experiment was not sufficient for in-depth analyses. This was mainly caused by continuous refinements of the methodology for the experiments over a long period of time and this delayed the final release of the experiments. In view of the slow pace of data gathering, a decision was made to limit the data collection to just one of the experiment types, i.e. experiment type "Drop Prerequisite

Skills". This meant that the initial set of experiments that were running be stopped. This decision was based on the fact that "Drop Prerequisite Skills" experiments had the most data collected that was useful.

While the data collection was ongoing, we realized that the initial design of this experiment introduces a possible attrition bias, mainly because the control group gets assigned more remediation assignments to complete than the experimental group, hence affecting their completion/drop-out rates (one of our outcome measures). [36, 89] As an example, suppose two students are selected for a given link in the graph. If one student is assigned to the experiment group, the remediation assignment associated with the prerequisite skill link will be marked as completed for this student. However, the student in the control group gets to be assigned that prerequisite skill, meaning that the students in the control group receive one additional assignment more than those in the treatment group. In analyzing this experiment, any differences between condition found could be attributed to the fact that one group had more work to do than the other, and not the effect of the treatment alone, threatening the internal validity of these experiments. In view of this obvious attrition bias, we redesigned the experiment type.

The fix this problem, we selected participants for this study only if there had remediation for at least two directly unrelated prerequisite skill links. We define two links to be directly unrelated if they do not share the same post-requisite skill. For illustrative purposes, suppose skill A is a prerequisite of skill B, and skill C is a prerequisite of an unrelated skill D. We will use both links in this study. Participants in this study were selected if they had need for remediation for all four skills from a single PLACEments test. Any participant in this student is considered in the control group for one of the links, in which case they are assigned the prerequisite skill for that link, and at the same time in the experimental group for the other link (in which case they are not assigned the prerequisite skill of that other link.) This way, participants in the study for each of the links will be assigned the same amount of work, reducing the possible attrition bias that resulted from the previous design. We therefore proceed to analyze the experiments based on this newly redesigned the student.

As shown in table 11-1 above, we use completion on the post-requisite skill as an important measure. We then perform an analysis of variance to determine the differences between the completion rates. We then use that to make statements about the links. We will proceed to analyze the two links for which we had collected enough data to analyze. We define "enough" as having at least 20 participants in both conditions.

11.5.1 Results of Drop Prerequisite Skill Experiment

We proceed to analyze the data for two of the about 170 links in the study. We selected these two links because they are the once with at least 20 participants whose data we have collected as of the time of this write-up. This experiment is particularly susceptible to low numbers of participants due to the fact that a very small percentage of students who use PLACEments have need of such a large set of remediation assignments after the assessment. The next subsections briefly describe the links and the results from the study related to those two links.

11.5.1.1 Adding Proper Fractions → Adding Mixed Numbers

Regarding this prerequisite skill link, the domain experts who designed the underlying PLACEments Skill Graph hypothesized that Adding Proper Fractions is a prerequisite to Adding Mixed numbers. These two are fifth-grade skills drawn from the 5.NF.A.1 standard in the Common Core State Standards for Mathematics. [17] Students who have gained these skills are expected to have the knowledge and ability to add proper fractions of the form $\frac{3}{4} + \frac{5}{6}$ and mixed numbers such as $4\frac{2}{7} + 7\frac{2}{5}$. Based on the hypothesis students should be able to add proper fractions before they can add mixed numbers.

For this link, there were only 91 participants, all of whom are middle school students who use the PLACEments feature in ASSISTments. Fifty of the students were randomly assigned to control and the other 41 assigned to treatment. To examine the research question associated with the "Drop Prerequisite Skill" experiment stated in the first row of table 11-1 above and for this link, we performed a chi-square test to understand the effect of condition on completion of the post requisite skill, and ANOVAs to understand condition's effect on mastery speed on the post-requisite skill and performance on the two initial items of the post-requisite assignment. To account for other measures of student prior performance on completion, we performed a series of logistic regressions in which we included these measures. The measures include prior start rate, prior completion rate and prior percent correct. Within the 91 students, the effect of condition on completion was found to be significant, $X^2(1, N=91) = 4.23$, p-value = 0.04. Analysis of the means revealed that students in the treatment condition (who were assigned the prerequisite skill) have a higher completion rate (M=0.26, SD=0.46, n=41) than those in the control condition (M=0.1, SD=0.30, n=50) who were not assigned the prerequisite skill. As a covariate, students' prior assignment completion rate was not significantly related to post-requisite skill completion though the p-

value seemed to lean towards significance (Model 2 of Table 11-2). Prior assignment start rate exhibited a similar behavior. (See model 1 of table 11-2). Model 4 in table 11-2 shows that the most important predictor of completion of the post-requisite skill's assignment is students' prior performance. When this feature is accounted for, the variance in completion contributed to by condition is borderline significant.

Table 11-2 Logistic regression of prior start and completion rates, performance and condition on completion of post-requisite skill assignment (Link 1)

Variable	*Model 1*	*Model 2*	*Model 3*	*Model 4*
Intercept	-1.46**	-1.30**	-1.44**	-8.26***
Condition (Drop Prereq)	-1.33*	-1.36*	-1.35*	-1.25+
Prior Start Rate	0.08+		0.87	3.53
Prior Completion Rate		1.44+	0.64	-3.86
Prior Performance				12.41***

+p<0.1; *p<0.05; **p<0.01; ***p<0.0001

Table 11-3 ANCOVA of the Effect of Condition on Speed of Mastery of Post-requisite skills (Accounting for average prior mastery speed)

Source	*df*	*SS*	*MS*	*F*	*p*
Prior Avg. Mastery Speed	1	3.45	3.45	0.482	0.274
Condition	1	9.06	9.06	1.263	0.496
Error	20	1.76	0.10		
Total	23	14.27			

To examine the effect of condition on the number of problems students complete to master the post-requisite skill (mastery speed), we observe that even after accounting for students' prior average completion rates, condition does not significant relate to relate to mastery speed, $F(1,23)=1.23$, $p=0.496$. This shows that the data we have collected so far does not support the hypothesis that being assigned practice of "Adding Proper Fractions" causes faster learning of the post-requisite skill, "Adding Mixed Numbers". See table 11-3 for the detailed results.

To further understand the effect of condition on the students' first two randomly selected items of the post-requisite assignment, we performed an ANCOVA, accounting for students' prior assignment start and completion rates. As table 11-4 shows, condition is not a significant predictor of performance on the two items in the post-requisite assignment, F(1,19)=1.46, p=0.242.

Table 11-4 ANCOVA of the effects of condition on post requisite performance (Link 1)

		Post-requisite Performance (First Two Items)			
Source	*df*	*SS*	*MS*	*F*	*p*
Prior Completion Rate	1	0.03	0.03	0.33	0.574
Prior Start Rate	1	0.14	0.14	1.43	0.247
Condition	1	0.15	0.15	1.46	0.242
Error	19	1.92	0.10		
Total	23	2.24			

11.5.1.2 Rounding to 10 and 100 → Rounding Whole Numbers

The second link which we report here is: Rounding to 10 and 100 → Rounding Whole Numbers. "Rounding to 10 and 100" is a third-grade skill from the standard 3.NBT.A.1 while "Rounding Whole Numbers" is a fourth-grade skill from the 4.NBT.A.3 standard. It is hypothesized that students need to learn to round numbers to either the tens or hundreds of places before they will be able to round numbers to the thousands and ten-thousands places.

To understand the effect of practice of the skill "Rounding to 10 and 100" on completion of the assignment related to the post-requisite skill, "Rounding Whole Numbers", we performed a chi-square test to analyze post-requisite assignment completion. We found that within the 21 students, the effect of condition on completion was not found to be statistically significant, $X^2(1, N=21) = 0.88$, p-value = 0.347. Controlling for the students' prior assignment start rate (Table 11.5), condition still did not have a significant relationship with assignment completion. Similar observations were made for the effect of condition on post-requisite assignment performance (i.e. performance on the two initial items).

Table 11-5 Logistic regression of prior start and completion rates, performance and condition on completion of post-requisite skill assignment (Link 2)

Variable	*Model 1*
Intercept	-1.46
Condition (Drop Prereq)	-1.33
Prior Start Rate	0.08
Prior Completion Rate	
Prior Performance	

$*p<0.05$; $**p<0.01$; $***p<0.0001$

Table 11-6 An ANOVA and ANCOVA of the effects of condition on mastery speed (accounting for prior mastery speed)

Source		Post-requisite Mastery Speed (condition only)				Post-requisite Mastery Speed (prior covariate)			
	df	*SS*	*MS*	*F*	*p*	*SS*	*MS*	*F*	*p*
Condition	1	12.85	12.85	2.367	0.142	12.85	12.85	2.264	0.152
Prior Mastery Speed						1.13	1.129	0.260	0.617
Error	17	92.31	5.43			90.83	5.677		
Total	19	105.16				104.81	19.66		

Table 11-7 ANCOVA of the Effects of Condition on Other Dependent Measures (Link 2)

Source		Post-requisite Performance (First Two Items)			
	df	*SS*	*MS*	*F*	*p*
Prior Completion Rate	1	0.03	0.03	0.53	0.479
Prior Start Rate	1	0.36	0.35	5.47	0.034
Condition	1	0.01	0.01	0.231	0.638
Error	15				
Total	19				

As shown in table 11-1, one of the outcome measures of interest is the speed with which students' complete assignment. We observe from Table 11-6 that regarding the second prerequisite link, condition does not

have any significant impact on the speed with which students learn the post-requisite skill $F(1,19)=2.37$ and $p=0.142$, even when we account for the students prior learning speed where $F(1,19)=2.264$ and $p=0.152$.

Additionally, for this current link, condition does not have any significant effect on other measures of student performance in the post-requisite skill. Table 11-6 evidently shows that accounting for prior students' assignment completion and start rates does not help the situation.

11.6 **Analyses of Results**

We set out to infer the relationship between skills incorporated in a domain-expert designed prerequisite skill structure graph. The graph consists of a number of prerequisite skill links that include the two which are described in section 11.4.1.1 and 11.4.1.2 above. We find from the results that when students get practice on the "Adding Proper Fractions" knowledge component, such students are significantly more likely to complete the assignment related to the post-requisite skill than if there are not, even after controlling for other prior performance measures. About 16% more of the students who are given practice on the pre-requisite skill are more likely to complete the post-requisite skill than those not assigned. This evidence seems to suggests that there is an existing prerequisite skill relationship between the two skills. Thus, we can conclude that "Adding Proper Fractions" is a prerequisite skill of "Adding Mixed Numbers."

The second skill link did not exhibit similar characteristics. As was seen from the results, we cannot conclude whether there exists a prerequisite skill relationship between "Rounding to 10 and 100" and "Rounding Whole Numbers" or not. In view of the limited data, we cannot make claims about the effect of condition on post-requisite skill performance. Students speed of learning the post-requsiute skill does not seem to be impacted by whether students are assigned to practice the "Rounding to 10 and 100" knowledge component or not. These findings may be due to the small amount of data that we collected, reducing the statistical power of our findings.

11.7 **Discussion**

As was stated earlier, learning progressions/prerequisite skill maps have informed the content related standards that have been developed and used by many school districts in the United States. These standards have been used as a basis for developing different frameworks for ordering content as teachers provide instructions to students.

While many of these frameworks have been used extensively, there is a lack of empirical data to support the effectiveness of these frameworks. This work reports initial baby steps that have been taken to use randomized experimentation to refine prerequisite skill structure graphs. While not all research questions about prerequisite skills were answered in this dissertation, we have shown that randomized controlled experimentation has promise in refining prerequisites skill graphs, augmenting the work of domain experts who design these frameworks and researchers who build student models to enhance student learning and develop interventions. Ultimately, students would benefit from teachers who use more accurate frameworks for teaching domain-specific content.

Additionally, several interventions have been proposed to solve many educational problems, particularly those that deal with students' inability to comprehend content. However, as noted by Ioannidis, a lot of research results are false [45]. Studies that relate to educational content may also be false because of the fact that the underlying framework from which students are taught may be influencing the many research findings. We proposed four experiment categories that PLACEments allows to be conducted, and we reported the results of one such experiment.

The evidence presented herein indicates that we are able to refine an existing prerequisite skill link just by running these randomized experiments. We showed that for one of the links, being assigned the prerequisite skill is essential for completion of assignments of the post-requisite skill. In the other link, the evidence we collected and analyzed was not sufficient to make any claims about the prerequisite skill link. Findings of this nature help in eliminating bad content ordering as one of the possible causes of false positives in numerous experiments that have been reported. Conversely, there are many educational-content related experiments that are not reported because they are perceived to have failed. This is usually referred to as publication bias [35]. Could it be that the interventions themselves may be effective, but the experiments have failed because the underlying progressions framework is not necessarily valid and hence the root of the experiment failure? Experiments of the nature presented in this dissertation could be helpful in reviving the failed studies, hence reducing some of the causes of failed experiments and subsequent publication bias.

Finally, students will ultimately benefit from progression frameworks that really achieve the goals intended by the developers of content-related standards like the CCSS. Additionally, if the findings indicate that the

stalling of student achievement in math for 4th grade students [60], this could lead to a positive rate of growth in the future.

11.8 **Limitations**

The experiments described and analyzed in this section provide a proof of concept that the features proposed in the new PLACEments infrastructure are helpful in finding portions or links in a prerequisite skill hierarchy that require more scrutiny. However, this design does not come with its own limitations, and we discuss one of these.

The major limitation of this work is the amount of time required for data collection. One of the actions that can be taken to address this limitation is to ensure that a data collection is orchestrated, i.e. a specific set of students should be selected based on some criteria, and a specified period of time given for the students/participants to take and complete predesigned placements tests and its accompanying remediation assignments. Predesigning the placements test for these experiments ensures that focus is given to a specific and usually small section of the prerequisite skill graph, instead of the entire graph as was done in the experiments described in this dissertation. While doing this, just one experiment should be run during that period. This will ensure that a large amount of data can be collected within a short period of time. This should give increase the power of any findings that may be made during the analysis of the experiments.

12 Does It really help to Assign Prerequisites Prior to Learning a new skill?

12.1 Introduction

Over the years, attempts have been made in the Educational Data Mining (EDM) community to find better ways to improve student learning. The approaches that have been studied vary drastically. While some have studied the medium of instruction as a means to improve student learning, others have investigated ways to improve methods of instruction, irrespective of the medium used. Still others have studied the emotional states of students during instruction and testing, with the goal of proposing effective interventions to improve student learning [11, 78] Among those who have studied the medium of instruction, several authors have focused on the use of intelligent tutoring systems (ITSs) as a means of improving student learning. Early on, Merrill, et al. [57] compared human tutors to intelligent tutoring systems with the aim of determining which was more effective. More recently, VanLehn [91] compared human tutors, intelligent tutoring systems, and other tutoring systems to determine their relative effectiveness at helping students learn.

Several ITSs (e.g., The Cognitive Tutor, ASSISTments) have been developed to improve student learning [7, 70]. ASSISTments uses tutoring strategies in the form of hints and scaffolds to help students solve difficult problems. As the name suggests, hints provide pointers to explanations that assist students in solving the problem under consideration. Alternatively, scaffolds offer a form of assistance that breaks the problem to be solved into smaller steps or segments which, when solved in the presented order, lead the student to a solution for the original problem. Each of these strategies can be initiated by the student. Multiple studies have been conducted within the ASSISTments system to compare the efficacy of these methods of assistance. After showing that ASSISTments led to student learning, Razzaq and Heffernan [72] investigated the effects of scaffolds and hints. Though not statistically reliable, they showed that scaffolds offered potentially better effects on learning than hints. However, the hints and scaffolds were based solely on the skill being assessed. In a follow-up study, Razzaq and Heffernan [71] looked at the amount of tutoring needed to assist students and found that the knowledge level of participants had a great impact on the amount and quality of tutoring required. Results suggested that low-knowledge students

benefited more from tutored problem solving than from sample solutions to problems, while high-knowledge students benefited more from seeing solutions than from tutored problem solving.

Whilst the above-mentioned and studied features seem to do well in assisting students to learn the skill under consideration, to the best of our knowledge, none of the previously reported studies have focused on the effects of prerequisite skills on student knowledge. It seems intuitive that knowledge of and ability to answer questions related to a skill, to a large extent, is dependent on the student's knowledge of the prerequisite skills[5] to the skill. In the current study, we hypothesize that an alternative way to assist students is to make them practice problems rooted in prerequisite skills. In other works, if students revisit these prerequisite skills and fill knowledge gaps, solving problems within the primary skill should be much quicker and easier. This hypothesis is stronger for skills that are a strict combination of their prerequisites. Moreover, we hypothesize that students need to know all the prerequisites of a skill to be able to learn and demonstrate he knowledge of the skill.

"Solving Equations" is a middle school math knowledge component that has been studied extensively by many math education and psychology experts. Rittle-Johnson and Star [76] report a study in which they compare three different methods in which students learning this knowledge components. They also studied the benefit of having prior knowledge of algebraic methods required to learn and solve equations and report that prior knowledge is an important predictor of performance in solving equations. [77] While this finding is interesting and appears to show the importance of prior knowledge, their definition of prior knowledge is a too broad. To the best of our knowledge, we have not found any research that has studied the effect of individual prerequisite skills of a given skill on performance of the skill. Moreover, based on the common core state standards [17], we find that the standard (8.EE.C.7b) from which this skill is drawn has four direct of prerequisite standards: Order of Operations, 1-Step Addition and Subtraction, 1-Step Multiplication and 1-Step Division. We are therefore hypothesizing that students should know all of these skills to be able to more easily learn the post-requisite skill.

12.2 **Research Questions**

The following research questions therefore arise from our hypotheses:

[5] See sections 2.1 and 2.2 for a complete definition of skills and prerequisite skills.

1. Does being assigned practice of all the prerequisite skills of "Solving Equations" impact completion of and performance on that skill?
2. Given that "Solving Equations" has four hypothesized prerequisite skills, does allowing students to choose which of the prerequisites to learn impact their performance on the skill?

12.3 **Methods**

12.3.1 Participants

Participants for this study were drawn from the population of ASSISTments [42] users. They include 390 students 6^{th} to 8^{th} grade students. These students were spread over 59 classes.

12.3.2 Experiment Design

The randomized experiment was designed to have three different conditions: "No Prerequisite Remediation", "Assign All Prerequisite Remediations" and "Student Choice of Remediation". Students in the study are assigned a pretest that has three "Equation Solving" problems. Participants who solve all three questions correctly are deemed to already know the skill and hence are dropped from the study.

As can be seen from Figure 12-1, students were randomly assigned to one of the three conditions stated above. Students assigned to the "No Remediation" condition are not given any practice on the prerequisite skills. They just get to practice more of the current skill. In the "Assign All Remediations" condition, students are assigned practice questions in all of the three remediation skills, after which they get to practice some questions in the post0requisite skill. In the third condition, students choose whether to practice all of prerequisite skills of "Solving Equations" or just a subset of their choice. Students in the choice condition either choose to "All Remediations" from the onset, or a subset of remediations. (Figure 12-2) All students were then given practice on the post-requisite skill. Thereafter, participants are presented with a 3-item a post-test that checks their ability to apply the knowledge gained to solve challenging questions of the post-requisite skill (Solving Equations).

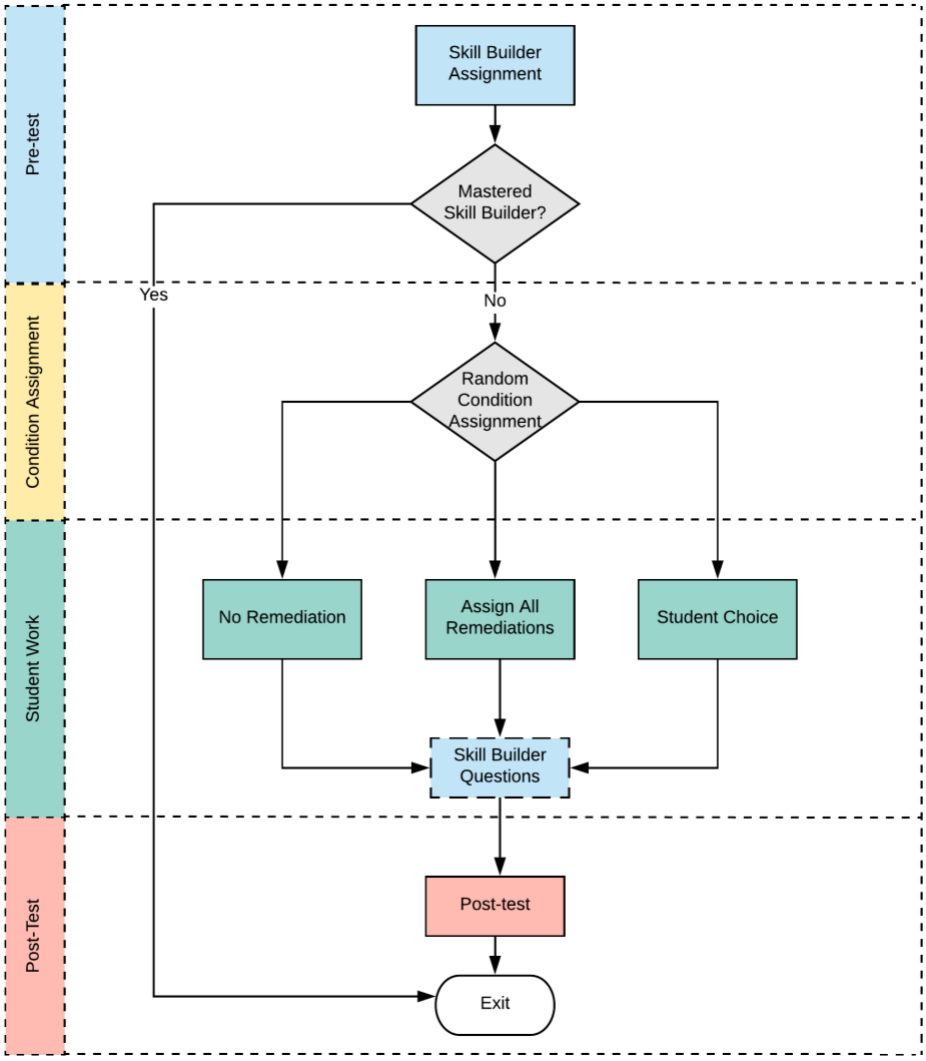


Figure 12-1 Experimental Design – Overall

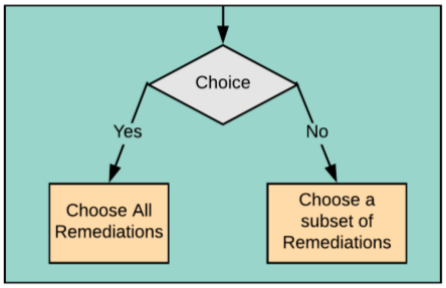


Figure 12-2 Options for Student Choice Condition

Students in both "Assign all Remediations" and "Student Choice" groups were informed about their need to practice additional remediations. Those in the "Assign all Remediations" are informed that they will be required to complete all the remediations before they can proceed. Figure 12-3 depicts the direction received by students in this condition. Students in the choice condition are given the same list of remediations as those in the "Assign All

Remediations" group and, are asked to either choose to complete all or practice a subset of the remediations. Figure 12-4 shows the prompt.

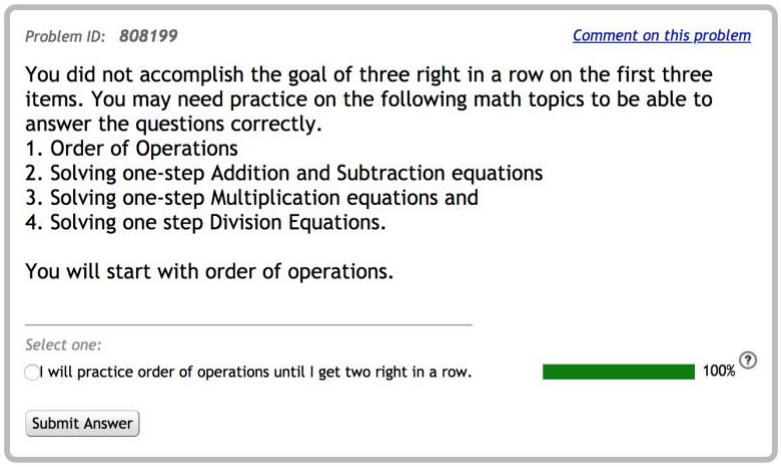


Figure 12-3 Prompt for Assign All Remediations

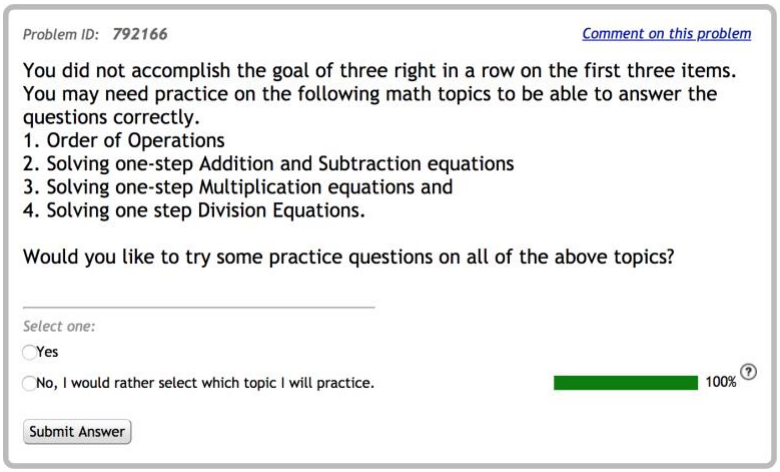


Figure 12-4 Prompt for "Student Choice"

12.4 **Random Assignment**

Prior to random assignment, students who demonstrated mastery of the initial skill by answering all three questions in the initial skill builder assignment without any errors were not assigned to any condition. We therefore excluded 140 students who fell in this group. Random assignment was performed at the student-level on the remaining 250 students. Of this number, 82 were randomized into "No Remediation", 79 to "Student Choice" and the remaining 89 to "Assign All Remediation."

12.5 **Results**

Among the students included in the experiment, all of whom made at least one error in the initial skill builder assignment, 52% made 1 error, 27 % made 2 errors and 21% made 3 errors. The mean score at pre-test was 1.57 out of three with a standard deviation of 0.83, showing a fair range of performance at pre-test/initial skill builder assignment.

To answer our initial research question, we performed an analysis of variance to determine how related the assignment to condition is to performance on the post-test, which was a more difficult set of questions of the post-requisite skill, measured by percent of students who got that problem correct. As was mentioned earlier, the post-test was composed of three more challenging items from the post-requisite skill. We proceed to analyze the remaining participants based on the following outcome measures: post-requisite completion and performance.

Attrition Rates

An analyses of variance (ANOVA) on completion yielded significant variance among conditions, $F(2,250)=8.9$, p-value = 0.01. A post hoc Tukey test showed that the "Assign All Remediations" group differed significantly from the "No Remediation" group, $p = 0.05$. Additionally, the same post-hoc analyses showed that the "Assign All Remediations" group and "Student Choice" groups also differed significantly, $p=0.02$. The "Student Choice" group was not significantly different from the group that was assigned no remediation. Table 12-1 shows the means and standard deviations for each experimental group.

Table 12-1 Completion Rates Per Condition

Condition	*n*	Completion		Post-test	
		Mean	SD	Mean	SD
No Remediation	82	0.80	0.40	0.7	0.39
Student Choice	79	0.82	0.38	0.69	0.38
Assign All Remediation	89	0.65	0.48	0.53	0.42

To better understand the choices made by students within the Student Choice condition, we performed a post-hoc analysis of the choice group. It must be noted that students chose either to complete all remediations or

to select a subset of the remediations. Of the 79 students in the choice condition, 37 opted to complete all remediations and the other 42 opted to learn a subset of the remediations whiles the remaining 42 opted to complete a subset of the 5 remediations ([0-5]). Table 12-2 summarizes the distribution of students and the choices they made.

Table 12-2 Post-hoc analyses of student choice group

Choice Option	Remediations (completed)	Student Count	Completion Rate	Mean Post-test Performance
Choose All	1	1	0.00	0.00(N/A)
	2	1	0.00	0.00(N/A)
	4	35	0.89	0.73(0.35)
Summary		***37***	***0.84***	***0.69***
Choose Subset	0	20	0.80	0.65(0.36)
	1	5	0.40	0.40(0.54)
	2	6	1.00	0.94(0.14)
	3	6	0.83	0.78(0.40)
	4	5	1.00	0.73(0.27)
Summary		***42***	***0.81***	***0.69***

We performed a chi-square test to understand the effect of remediation choice option (i.e. Choose All and Choose Subset) on completion and find no significant difference between students who, from the onset, choose to complete all the assignments and those who choose to complete a subset of the remediations ($F(1,79)=0.11$, $p>0.05$). For the "Choose Subset" group, we built a simple logistic regression model using only the number of remediations started and/or completed as independent variable and the completion of the post-test as the dependent variable. We find that the number of remediations chosen significantly predicts whether a student completes the assignment or not with p-value < 0.01. This result is a bit surprising, particularly since the overall study showed no significant difference between students in the "No Remediation" and the "Student Choice" conditions. This seems to suggest that when students are given the choice, they make choices that seem to help them complete remedial assignments better than if they are forced to complete all the remedial assignments. As Table 12-2 suggests, students who chose not to complete any of the remediations had a high completion rate, but the least performance on post-test. The table further shows that students who chose to practice all the remediations (either from the onset or later) have high completion rates and high post-test performance. Though a

possible bias is observed from Table 12-2, the information presented seems to be suggesting that students must make wise choices to be able to gain the most benefit from the material.

Post-test (Knowledge Transfer)

In order to analyze post-test performance, all students in each of the conditions who attrited were assigned a post-test score of zero in order to ensure a fair comparison among the conditions. A trend similar to the observations regarding attrition is observed for post-test performance. ANOVA of post-test scores of students indicates a significant variance amongst conditions as well, $F(2,250)=5.27$, $p<0.01$. Another post-hoc Tukey test showed a significant difference between the "No Remediation" and "Assign All Remediation" groups, $p=0.01$; and between "Assign All Remediation" and "Student Choice" groups, $p=0.02$. There was however no significant difference between the "No Remediation" group and the Choose Remediation group. Table 12-3 summarizes the analyses of variance that was conducted for post-test score.

Table 12-3 ANOVA of the Effect of Condition Post-test Performance

		Post-Test Performance			
Source	*df*	*SS*	*MS*	*F*	*p*
Condition	2	1.65	0.83	5.27	0.01
Error	247	38.80	0.16		
Total	250	40.45			

12.6 **Implications**

The results suggest that requiring students to complete all remedial assignments prior to a classroom activity could be detrimental to the learning process. This could be a result of several factors including the case where students may not be completing the subsequent assignments due to the fact that they could be spending too much time on the remedial assignments, to the detriment of learning the post-requisite skill. As the results suggest (see Table 12-1) students who are forced to complete all required remedial assignments significantly perform worse than their counterparts who are either asked to make choices of which remedial skills to cover or who are not assigned remedial assignments at all.

The results of this study do seem contradictory to the widely held notion that students need to know the prerequisite skills to be able to learn the post-requisite skills. In fact, several studies have shown that when students are given remedial math classes prior to taking college-level courses, they are significantly less likely to do well in the college-level course which is supposedly the post-requisite of the remedial classes.[54] In fact there are other findings that show that remedial courses at the college level may not be good for all kinds of students [9], and that it is what and how the remedial courses are taught that may contribute to students' higher achievement in the college-level course. [22] The findings of this current study may suggest a similar set of characteristics showing up at the elementary and middle school levels of education. Further analyses of these findings are required to be able to draw conclusions about relationships between these skills, particularly since this present study only focused on one skill.

12.7 **Conclusion**

We set out to investigate the effect of prerequisite skill remediation on post-requisite skills. The ultimate goal for this study was to understand the effect of prerequisites on student knowledge. In other words, does it matter to assign students practice assignments of prerequisites of a new knowledge component they are about to learn. In particular, we wanted to find out whether assigning students remediations or forcing them to choose remediations improves student achievement in the subsequent post-requisite skill's assignment. Using a randomized controlled trial, which had three variations of remedial assignments, we have shown that assigning students remediations could be detrimental to their learning the post-requisite skill. However, allowing students to make choices of which prerequisites to learn prior to learning a new skill may be better than requiring them complete all prerequisites.

Like all other studies, this study is not without its limitations. First and foremost, the skill chosen may have had an impact on the differential dropout rates across condition. Since "Equation Solving" is a skill that has four different prerequisite skills, it may be that students who were assigned to the complete all remediations assignments group may have been assigned an unfair amount of work, compared with the other groups. Additionally, the order of the prerequisite skills was imposed on the students. A better design will just be to allow students in both "Student Choice" and "Assign All Remediations" to complete the remediations in the order of the

students' choice. Future work will investigate the effect of order on students' completion of remediations and subsequent post-test.

13 Conclusion

Prerequisite sill graphs have been used for many decades to represent the order in which content is expected to be taught in order to improve student's achievement in state exams and assist them in the easier comprehension of content. While these models have been largely designed by domain experts based on content-related literature, there has been a growing desire to find ways to use data-driven methods to refine these domain-expert-designed models. As the work presented in this dissertation shows, data mining techniques have been used with promising results.

In this work, we have used additional data mining techniques to refine learning maps, with varying degrees of success. We also have presented an adaptive testing system, PLACEments, that uses domain-expert-designed prerequisite skill structures for assessment of student knowledge. This system additionally offers remedial practice to help low-achieving students fill any knowledge gaps identified during assessment. We have shown that this infrastructure presents a useful platform for using randomized controlled experiments to refining portions of these domain-expert-designed graphs. We presented examples of six types of experiments, one of which have been run in PLACEments, and the possible causal statements that can be made about the links in the underlying prerequisite skill structure.

While the method presented in this dissertation has some promise and has achieved moderate success, it comes with a number of limitations. The first of which is the time it takes to collect substantial real time data for each of the experiments. The second is the limited amount of data that can be collected per link during the period of the experimentation. This limitation becomes apparent when the size of the underlying prerequisite skill structure is large, i.e. the number of prerequisite skill links to be studied is large. Also, since the setup for the experiments require the inputs of teachers, the set of links that get used in the experiments is dictated by the selections made by teachers as they assign these tests. One proposed solution to these limitations is to control the selection of participants as well as skill links to be used. By this we suggest pre-selecting focused subsets of skill links to examine through experimentation and administer predesigned PLACEments tests encompassing those specific links of interest to a population that is representative of students identified to need practice within such skills. There could be several other proposals that we can employ to fix the issues of low numbers of participants.

14 References

1. Adjei, S., et al. *Refining learning maps with data fitting techniques: Searching for better fitting learning maps*. in *Educational Data Mining 2014*. 2014.

2. Adjei, S., et al., *Refining Learning Maps with Data Fitting Techniques: Searching for Better Fitting Learning Maps*., Z.P. J. Stamper, M. Mavrikis, & B. M. McLaren (Eds.), Editor. 2014: 7th International Conference on Educational Data Mining p. 413-414.

3. Adjei, S.A., A.F. Botelho, and N.T. Heffernan, *Predicting student performance on post-requisite skills using prerequisite skill data: an alternative method for refining prerequisite skill structures*. 2016. p. 469-473.

4. Adjei, S.A. and N.T. Heffernan, *Improving Learning Maps Using an Adaptive Testing System: PLACEments*, in *Artificial Intelligence in Education*, C. Conati, et al., Editors. 2015, Springer International Publishing %8 2015-01-01. p. 517-520.

5. Adjei, S.A. and N.T. Heffernan, *Improving Learning Maps Using an Adaptive Testing System: PLACEments*, in *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings*, C. Conati, et al., Editors. 2015, Springer International Publishing: Cham. p. 517-520.

6. Amabile, T.M. and J. Gitomer, *Children's artistic creativity: Effects of choice in task materials.* Personality and Social Psychology Bulletin, 1984. **10**(2): p. 209-215.

7. Anderson, J.R., et al., *Cognitive tutors: Lessons learned.* The journal of the learning sciences, 1995. **4**(2): p. 167-207.

8. Anderson, J.R. and C. Lebiere, *The atomic components of thought*. 1998, Mahwah, N.J: Lawrence Erlbaum Associates.

9. Angela, B. and L. Bridget Terry, *Does Remediation Work for All Students? How the Effects of Postsecondary Remedial and Developmental Courses Vary by Level of Academic Preparation.* Educational Evaluation and Policy Analysis, 2017. **40**(1): p. 29-58.

10. Ashcraft, M.H. and E.P. Kirk, *The relationships among working memory, math anxiety, and performance.* Journal of Experimental Psychology: General, 2001. **130**(2): p. 224-237.

11. Baker, R.S., et al., *Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments.* International Journal of Human-Computer Studies, 2010. **68**(4): p. 223-241.

12. Barnes, T. *The q-matrix method: Mining student response data for knowledge*. in *Educational Data Mining Workshop*. 2005.

13. Block, J.H. and R.B. Burns, *Mastery Learning.* Review of Research in Education, 1976. **4**: p. 3-49.

14. Botelho, A., H. Wan, and N. Heffernan. *The prediction of student first response using prerequisite skills*. in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. 2015. ACM.

15. Broaddus, A., A. Sharma, and S. Adjei, *Using Test Responses to Validate a Learning Map of Integer Understanding.*

16. Brunskill, E. *Estimating Prerequisite Structure From Noisy Data*. 2011. Citeseer.

17. CCSS-MA, *Common Core State Standards for Mathematics.*, N.G.A.C.f.B.P.a.t.C.o.C.S.S. Officers, Editor. 2010: Washington, DC.

18. Cen, H., *Generalized learning factors analysis: improving cognitive models with machine learning*. 2009: Carnegie Mellon University.

19. Cen, H., K. Koedinger, and B. Junker. *Learning factors analysis–a general method for cognitive model evaluation and improvement*. 2005. Springer.

20. Chaplot, D. and K.R. Koedinger, *Data-driven Automated Induction of Prerequisite Structure Graphs.* 2016.

21. Chen, Y., P.-H. Wuilemin, and J.-M. Labat, *Discovering Prerequisite Structure of Skills through Probabilistic Association Rules Mining*, in *8th International Conference on Educational Data Mining*, J.G.B. Olga C. Santos, Cristobal Romero, Mykola Pechenizkiy, Agathe Merceron, Piotr Mitros, José María Luna, Cristian Mihaescu, Pablo Moreno, Arnon Hershkovitz, Sebastian Ventura, and Michel Desmarais, Editor. 2015: Madrid, Spain. p. 117 - 125.

22. Christopher, L.Q. and D. Mickey, *Is Learning in Developmental Math Associated With Community College Outcomes?* Community College Review, 2016. **45**(1): p. 33-51.

23. Clement, J., *Overcoming students' misconceptions in physics: fundamental change in children's physics knowledge.* Journal of research in science teaching, 1987. **28**: p. 785.

24. Clements, D.H. and J. Sarama, *Learning trajectories in mathematics education.* Mathematical thinking and learning, 2004. **6**(2): p. 81-89.

25. Confrey, J., A.P. Maloney, and A.K. Corley, *Learning trajectories: a framework for connecting standards with curriculum.* ZDM, 2014. **46**(5): p. 719-733.

26. Corbett, A.T. and J.R. Anderson, *Knowledge tracing: Modeling the acquisition of procedural knowledge.* User modeling and user-adapted interaction, 1994. **4**(4): p. 253-278.

27. Cordova, D.I. and M.R. Lepper, *Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice.* Journal of educational psychology, 1996. **88**(4): p. 715.

28. Desmarais, M., X. Pu, and J.G. Blais. *Partial Order Knowledge Structures for CAT Applications*. in *GMAC Conference on Computerized Adaptive Testing*. 2007.

29. Desmarais, M.C., B. Beheshti, and R. Naceur, *Item to Skills Mapping: Deriving a Conjunctive Q-matrix from Data*, in *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings*, S.A. Cerri, et al., Editors. 2012, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 454-463.

30. Desmarais, M.C. and M. Gagnon. *Bayesian student models based on item to item knowledge structures*. in *European Conference on Technology Enhanced Learning*. 2006. Springer.

31. Desmarais, M.C., A. Maluf, and J. Liu, *User-expertise modeling with empirically derived probabilistic implication networks.* User modeling and user-adapted interaction, 1995. **5**(3-4): p. 283-315.

32. Desmarais, M.C., P. Xu, and B. Beheshti, *Combining techniques to refine item to skills Q-matrices with a partition tree.* International Educational Data Mining, 2015.

33. Dogan-Dunlap, H., *Lack of set theory relevant prerequisite knowledge.* International Journal of Mathematical Education in Science and Technology, 2006. **37**(4): p. 401-410.

34. Doignon, J.-P., et al., *Knowledge spaces*. 1 ed. 1999, New York;Berlin;: Springer.

35. Egger, M. and G.D. Smith, *Bias in location and selection of studies.* BMJ: British Medical Journal, 1998. **316**(7124): p. 61.

36. Ellenberg, J.H., *Selection bias in observational and experimental studies.* Statistics in medicine, 1994. **13**(5-7): p. 557-567.

37. Embretson, S.E., *A cognitive design system approach to generating valid tests: Application to abstract reasoning.* Psychological Methods, 1998. **3**(3): p. 380.

38. Flowerday, T. and G. Schraw, *Effect of Choice on Cognitive and Affective Engagement.* The Journal of Educational Research, 2003. **96**(4): p. 207-215.

39. Friedman, N., I. Nachman, and D. Peér. *Learning bayesian network structure from massive datasets: the «sparse candidate «algorithm*. in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. 1999. Morgan Kaufmann Publishers Inc.

40. Gagne, R.M., *Presidential address of division 15 learning hierarchies.* Educational psychologist, 1968. **6**(1): p. 1-9.

41. Gierl, M.J., C. Wang, and J. Zhou, *Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT.* Journal of Technology, Learning, and Assessment, 2008. **6**(6): p. n6.

42. Heffernan, N.T. and C.L. Heffernan, *The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching.* International Journal of Artificial Intelligence in Education, 2014. **24**(4): p. 470-497.

43. Hochreiter, S., *The vanishing gradient problem during learning recurrent neural nets and problem solutions.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998. **6**(02): p. 107-116.

44. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

45. Ioannidis, J.P.A., *Why most published research findings are false.* PLoS Med, 2005. **2**(8): p. e124.

46. Jeroen, J.G.v.M. and J. Sweller, *Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions.* Educational Psychology Review, 2005. **17**(2): p. 147-177.

47. Kaptchuk, T.J., *The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf?* Journal of Clinical Epidemiology, 2001. **54**(6): p. 541-549.

48. Khajah, M., R.V. Lindsey, and M.C. Mozer, *How deep is knowledge tracing?* arXiv preprint arXiv:1604.02416, 2016.

49. Koedinger, K.R., A.T. Corbett, and C. Perfetti, *The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning.* Cognitive Science, 2012. **36**(5): p. 757-798.

50. Larger, E. and J. Rodin, *The effects of choice and enhanced personal responsibility for the aged.* Journal of Personality and Social Psychology, 1976. **34**: p. 191-198.

51. Leighton, J.P., M.J. Gierl, and S.M. Hunka, *The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach.* Journal of educational measurement, 2004. **41**(3): p. 205-237.

52. Leszczenski, J.M. and J.E. Beck. *What's in a word? Extending learning factors analysis to model reading transfer*. in *13th International Conference on Artificial Intelligence in Education, Educational Data Mining Workshop*. 2007.

53. Li, N., et al. *A machine learning approach for automatic student model discovery*. in *Educational Data Mining 2011*. 2010.

54. Logue, A.W., W.-R. Mari, and D. Daniel, *Should Students Assessed as Needing Remedial Mathematics Take College-Level Quantitative Courses Instead? A Randomized Controlled Trial.* Educational Evaluation and Policy Analysis, 2016. **38**(3): p. 578-598.

55. Mahigir, F. and A. Karimi, *Parents socio economic background, mathematics anxiety and academic achievement.* International Journal of Educational Administration and Policy Studies, 2012. **4**(8): p. 177-180.

56. Meece, J.L., A. Wigfield, and J.S. Eccles, *Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics.* Journal of Educational Psychology, 1990. **82**(1): p. 60-70.

57. Merrill, D.C., et al., *Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems.* The Journal of the Learning Sciences, 1992. **2**(3): p. 277-305.

58. Mislevy, R.J., L.S. Steinberg, and R.G. Almond, *Design and analysis in task-based language assessment.* Language testing, 2002. **19**(4): p. 477-496.

59. Murayama, K., et al., *Don't Aim Too High for Your Kids: Parental Overaspiration Undermines Students' Learning in Mathematics.* J Pers Soc Psychol, 2015.

60. NAEP. *2015 Mathematics & Reading Assessments*. 2015 [cited 2016 February 20, 2016]; Available from: http://www.nationsreportcard.gov/reading_math_2015/#mathematics/scores?grade=4.

61. Osborne, J.W. and A. Overbay, *The power of outliers (and why researchers should always check for them).* Practical assessment, research & evaluation, 2004. **9**(6): p. 1-12.

62. Ostrow, K.S. and N.T. Heffernan. *The role of student choice within adaptive tutoring*. in *International Conference on Artificial Intelligence in Education*. 2015. Springer.

63. Pavlik Jr, P.I., H. Cen, and K.R. Koedinger, *Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models.* Online Submission, 2009.

64. Pavlik Jr, P.I., et al., *Using Item-Type Performance Covariance to Improve the Skill Model of an Existing Tutor.* Online Submission, 2008.

65. Pavlik, P.I., H. Cen, and K.R. Koedinger, *Performance Factors Analysis --A New Alternative to Knowledge Tracing*, in *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*. 2009, IOS Press. p. 531-538.

66. Perlmuter, L.C. and R.A. Monty, *The importance of perceived control: Fact or fantasy? Experiments with both humans and animals indicate that the mere illusion of control significantly improves performance in a variety of situations.* American Scientist, 1977. **65**(6): p. 759-765.

67. Piech, C., et al., *Deep knowledge tracing*, in *Advances in Neural Information Processing Systems 28*, C. Cortes, et al., Editors. 2015, Curran Associates, Inc. p. 505-513.

68. Popham, W.J., *Transformative assessment in action: An inside look at applying the process*. 2011: ASCD.

69. Rainey, R.G., *The effects of directed versus non-directed laboratory work on high school chemistry achievement.* Journal of Research in Science Teaching, 1965. **3**(4): p. 286-292.

70. Razzaq, L., et al. *The Assistment project: Blending assessment and assisting*. 2005.

71. Razzaq, L. and N.T. Heffernan. *Scaffolding vs. hints in the Assistment System*. Springer.

72. Razzaq, L. and N.T. Heffernan. *Scaffolding vs. Hints in the Assistment System*. in *Intelligent Tutoring Systems*. 2006. Berlin, Heidelberg: Springer Berlin Heidelberg.

73. Reigeluth, C.M., *Order, first step to mastery: An introduction to sequencing in instructional design.* In order to learn: How the sequence of topics influences learning, 2007. **2**: p. 19.

74. Rigney, J.W., *Learning strategies: A theoretical perspective.* Learning strategies, 1978. **165**.

75. Ritter, S., et al. *How Mastery Learning Works at Scale*. 2016. ACM.

76. Rittle-Johnson, B. and J.R. Star, *Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving.* Journal of Educational Psychology, 2009. **101**(3): p. 529-544.

77. Rittle-Johnson, B., J.R. Star, and K. Durkin, *The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving.* Journal of Educational Psychology, 2009. **101**(4): p. 836-852.

78. Robison, J., S. McQuiggan, and J. Lester. *Evaluating the consequences of affective feedback in intelligent tutoring systems*. IEEE.

79. Roschelle, J., et al., *Online Mathematics Homework Increases Student Achievement.* AERA Open, 2016. **2**(4).

80. Rose, D. and J. Gravel, *Universal design for learning guidelines: Version 2.* Wakefield, MA: National Centre on Universal Design for Learning, 2011.

81. Ruiz-Primo, M.A. and R.J. Shavelson, *Problems and issues in the use of concept maps in science assessment.* Journal of research in science teaching, 1996. **33**(6): p. 569-600.

82. Scheines, R., E. Silver, and I. Goldin. *Discovering prerequisite relationships among knowledge components*. in *Proceedings of the 7th International Conference on Educational Data Mining*. 2014.

83. Shapka, J.D., J.F. Domene, and D.P. Keating, *Trajectories of career aspirations through adolescence and young adulthood: Early math achievement as a critical filter.* Educational Research and Evaluation, 2006. **12**(4): p. 347-358.

84. Sheehan, K.M., *Tree-based approach to proficiency scaling and diagnostic assessment*. 2000, Google Patents.

85. Simon, M.A., *Reconstructing Mathematics Pedagogy from a Constructivist Perspective.* Journal for Research in Mathematics Education, 1995. **26**(2): p. 114-145.

86. Tatsuoka, K.K., *Rule space: An approach for dealing with misconceptions based on item response theory.* Journal of Educational Measurement, 1983. **20**(4): p. 345-354.

87. Tatsuoka, K.K., *Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach.* Cognitively diagnostic assessment, 1995: p. 327-359.

88. Torgerson, C.J. and D.J. Torgerson, *The Need for Randomised Controlled Trials in Educational Research.* British Journal of Educational Studies, 2001. **49**(3): p. 316-328.

89. Torgerson, D.J. and C.J. Torgerson, *Avoiding bias in randomised controlled trials in educational research.* British journal of educational studies, 2003. **51**(1): p. 36-45.

90. Van de Walle, J.A., et al., *Teaching Student-Centered Mathematics: Developmentally Appropriate Instruction for Grades 6 - 8*. Vol. 1. 2014: Pearson.

91. VanLehn, K., *The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems.* Educational Psychologist, 2011. **46**(4): p. 197-221.

92. Vuong, A., T. Nixon, and B. Towle. *A Method for Finding Prerequisites Within a Curriculum*.

93. Wan, H. and J.B. Beck, *Considering the Influence of Prerequisite Performance on Wheel Spinning.* International Educational Data Mining Society, 2015.

94. Wang, M.C. and B. Stiles, *An investigation of children's concept of self-responsibility for their school learning.* American Educational Research Journal, 1976. **13**(3): p. 159-179.

95. Wang, Y.-Z., *A GA-based methodology to determine an optimal curriculum for schools.* Expert Systems with Applications, 2005. **28**(1): p. 163-174.

96. Watanabe, M. and P. Sturmey, *The Effect of Choice-Making Opportunities During Activity Schedules on Task Engagement of Adults with Autism.* Journal of Autism and Developmental Disorders, 2003. **33**(5): p. 535-538.

97. West, S.G., et al., *Alternatives to the Randomized Controlled Trial.* American Journal of Public Health, 2008. **98**(8): p. 1359-1366.

98. Witzel, B.S. and P.J. Riccomini, *Optimizing Math Curriculum to Meet the Learning Needs of Students.* Preventing School Failure: Alternative Education for Children and Youth, 2007. **52**(1): p. 13-18.

99. Xu, Y. and J. Mostow. *Logistic Regression in a Dynamic Bayes Net Models Multiple Subskills Better!* in *Educational Data Mining 2011*. 2010.

100. Zhang, J. and I. King, *Topological Order Discovery via Deep Knowledge Tracing*, in *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part IV*, A. Hirose, et al., Editors. 2016, Springer International Publishing: Cham. p. 112-119.

101. Zuckerman, M., et al., *On the importance of self-determination for intrinsically-motivated behavior.* Personality and Social Psychology Bulletin, 1978. **4**(3): p. 443-446.