Towards Assessing Students' Fine Grained Knowledge: Using an Intelligent Tutor for Assessment

by

Mingyu Feng

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

by

August 18, 2009

APPROVED:

Professor Neil T. Heffernan Advisor	Professor Carolina Ruiz Committee Member
Professor Joseph E. Beck Committee Member	Professor Kenneth R. Koedinger Carnegie Mellon University External Committee Member
Professor Michael Gennert Head of Department	

Abstract

Secondary teachers across the United States are being asked to use formative assessment data to inform their classroom instruction. At the same time, critics of US government's No Child Left Behind legislation are calling the bill "No Child Left Untested". Among other things, critics point out that every hour spent assessing students is an hour lost from instruction. But, does it have to be? What if we better integrated assessment into classroom instruction and allowed students to learn during the test? This dissertation emphasizes using the intelligent tutoring system as an assessment system that just so happens to provide instructional assistance during the test.

Usually it is believed that assessment get harder if students are allowed to learn during the test, as it's then like trying to hit a moving target. So, my results are somewhat shocking that by providing tutoring to students while they are assessed I actually improve the assessment of students' knowledge.

Most traditional assessments treat all questions on the test as sampling a single underlying knowledge component. Yet, teachers want detailed, diagnostic reports to inform their instruction. Can we have our cake and eat it, too? In this dissertation, I provide solid evidence that a fine-grained skill model is able to

predict state test scores better than coarser-rained models, as well as being used to give teachers more informative feedback that they can reflect on to improve their instruction.

The contribution of the dissertation lies in that it established novel assessment methods to better assess students in intelligent tutoring systems. Through analyzing data of more than 1,000 students across two years, it provides strong evidence implying that it is possible to develop a continuous assessment system that can do all three of these things at the same time: 1) accurately and longitudinally assesses students, 2) gives fine grained feedback that is more cognitively diagnostic, and 3) saves classroom instruction time by assessing students while they are getting tutoring.

Acknowledgements

This dissertation and the growth in my knowledge over the last few years owe a great deal to many professors, colleagues, and friends. First among them is my advisor, Neil. He always encourages me to use the first person singular pronoun in dissertation, but there is no true way to separate out the contribution that he made to this dissertation. He inspired my interests in intelligent tutoring system and educational research and gave me direction by suggesting interesting problems. His valuable advice, encouragement, trusts have always been an inspiration. I also thank deeply for his generous sponsorship throughout my whole Ph.D. studies. I feel fortunate to have him and his wife, Cris, being good friends of mine. Their tips and help have made my life in the United States more enjoyable.

My thanks go to the members of my Ph.D. committee, Carolina, Joe and Ken, who provided valuable feedback and suggestions to my comprehensive-exam, my dissertation proposal talk and dissertation drafts. All these helped to improve the presentation and contents of this dissertation. I thank Ken and Joe for discussing

research ideas with me, and put into time and efforts with me in several papers that we published together. I truly appreciate that.

I thank all previous and current members of the ASSISTment project for their hard work on building the system that we share as a team, including Leena, Sanket, Tom, Beth, Goss, Terrence, the Michaels, Jason, Ruta, Jozsef, Zach, Abe, Yu, Shane, and all others who have contributed to the ASSISTments system. Leena has been an excellent companion in the past 5 years. I thank her for editing my paper to make it English. I thank Zach for his valuable discussions and co-working on my dissertation task of evaluating skill models of various granularities. I also want to acknowledge Hao Cen for his help on LFA.

Finally, I would like to thank my family. My husband, Song, gave me his understanding and love during the past few years. His support and encouragement is ultimately what made this dissertation possible. My parents receive my deepest gratitude and love for their dedication. Their patience with me while in graduate school is much appreciated. My mom must be happy to know that I am going to enter the real world finally. Special thanks also go to my little daughter Allison (You-ran), her nice cooperation in the past year made me possible to complete my studies on schedule.

Ming (Aug, 2009)

Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

1.1 The Problems

Across the world there are interests in student learning and assessing. In many US states there are also concerns about poor student performance on new high-stakes standards based tests that are required by *No Child Left Behind Act* (NCLB, 2002) legislation of the US government. To address this issue, educational technologies, like intelligent tutoring systems (ITS) have been developed and proven to be useful to help students learn. For instance, the Cognitive Tutors has been shown to lead to large learning gains (Koedinger et al, 1997). Recently, U.S. President Obama made a commitment to increase investment on educational software, saying that "[W]e will devote more than 3 percent of our GDP to research and development. … Just think what this will allow us to accomplish: solar cells as cheap as paint, and green buildings that

produce all of the energy they consume; *learning software as effective as a personal tutor*…" [1]. This sounds very inspiring. But, how do we build ITSs that are as effective as a personal tutor? Among other things, knowledge intensive student modeling is, in my view, a critical aspect for the success of any ITS. Assessment of an examinee's ability is the first step of student modeling in an ITS because student state is a prerequisite for creating a pedagogical strategy. The student model provides valuable information for the tutor to help on its tutoring strategy (e.g. when to interrupt and what to say when interrupt), problem sequencing (e.g. what's the next appropriate task to give to a student so his learning gain is maximized), performance prediction (how a student will respond to a step associated with certain rules), etc. It can be very useful for informing real teachers' of student's weakness so that they can adjust classroom instruction as well.

Assessment (or tests) has a long history of being used for performance evaluation. Yet, nowadays, we see both challenges and opportunities for assessment. On the one hand, the movement towards high-stakes testing promises to encourage rigor and accountability in middle school mathematics. But, as tests become more and more intensively employed, new concerns about over-testing arise. On the other hand, educational software has goes to schools and produced

[1] http://my.barackobama.com/page/community/post/amyhamblin/gGxW3n.

enormous learning (Koedinger et al., 1997). Comes after that is the question of how to assess students accurately in such a learning intensive environment, and, meanwhile, to respond effectively to teachers' needs for classroom instruction.

1.1.1 Formative assessment steals time from instruction

Across the country of the United States, high-stakes standards-based tests are required by the NCLB. The new testing system in Massachusetts, Massachusetts Comprehensive Assessment System (MCAS) requires students to take rigorous tests in English, math, history and science in grades 3-12. Currently, students need to pass the math and English portions of the 10th grade versions in order to get a high school diploma. However, as reported by Massachusetts Dept. of Education, in the year of 2003, 20% of students failed the 10th grade math test on the first try; in urban districts, a staggering 18% of students that take the 10th grade math MCAS never manage to pass. The industrial city of Worcester is a representative of many such districts. Moreover, the state of Massachusetts singled out student performance on the 8th grade math test as an area of highest need for improvement[2]. Partly in response to this pressure, and partly because teachers, parents, and other stakeholders want and need more immediate feedback about how students are doing, there has recently been intense interest in "Formative

[2] http://www.doe.mass.edu/mcas/2002/results/summary.pdf

Assessment" in K-12 Education (Boston, 2002; Black & Wiliam, 1998a, 1998b; Roediger & Karpicke, 2006) and in predicting student performance on end-of-year tests (Olson, 2005) with many companies[3] providing such services of practice tests. Some teachers make extensive use of these practice tests and released test problems to help identify learning deficits for individual students and the class as a whole. However, such formative assessments not only require great effort and dedication, but they also take valuable time away from instruction. Critics of No Child Left Behind are calling the bill "No Child Left Untested" emphasizing the negative side of assessment, in that every hour spent assessing students is an hour lost from instruction. Some online testing systems (such as Renaissance Learning, www.renlearn.com) automatically grade students and provide reports but they may not be informative as they do not maintain sufficiently rich data records for students and hence cannot report on a fine-grained level of student knowledge.

The limited classroom time available in middle school mathematics classes has always compelled teachers to choose between time spent assisting students' development and time spent assessing students' abilities. Therefore, a solution must involve a way whereby students can take an assessment, but also learn as

[3] Including Northwest Evaluation Association (http://nwea.org/assessments/), Measured Progress (http://measuredprogress.org), Pearson Assessments (http://www.pearsonassessments.com/). and the Center for Data-Driven Reform in Education (http://www.cddre.org/)

they are being assessed. A solution needs to be found so that teachers can get the benefit of being data-driven in trying to meet instructional objectives, but at the same time, make sure that their students' time is spent primarily on learning. The goal of integrating assessment and assisting instruction was achieved in the ASSISTments system that I will describe later.

1.1.2 Teachers' needs have been forgotten

Along with the change towards high-stakes testing is the fundamental dilemma that teachers face in trying to use assessment to guide instruction. Because assessment takes time away from instruction, how can teachers be sure that the time they spent on assessing will improve instruction enough to justify the cost of the lost instructional time? Worcester Public Schools are representatives of the many districts that are trying to address the dilemma. Among other things, they seek to use the MCAS assessments in a data-driven manner to provide regular and ongoing feedback to teachers and students on progress towards instructional objectives. For instance, the School Improvement Teams at each school review the results from the previous year to analyze which problems their students performed particularly poorly on. However, teachers want feedback much more often than once a year, but it usually takes the state six months to grade the MCAS exams. Moreover, they want better feedback than they currently receive. While the number of mathematics skills and concepts that a student needs

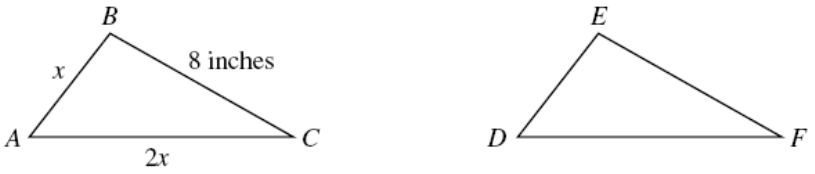
to acquire is on the order of hundreds, the feedback on the MCAS to principals, teachers, parents, and students is broken down into only 5 mathematical reporting categories, known as "Strands." (See Figure 1.1) In fact, the state's Curriculum Framework breaks the 5 strands into 39 individual "learning standards" for 8th grade math and tags each item with one of the 39 standards. The MCAS reporting system is a representative of other states' reporting system. In 2004, a principal handed us a report (shown Figure 1.1) he received from the state and asked that we focus efforts on Geometry and Measurement because his students scored poorly in those areas (receiving 38% and 36% correct compared to over 41+% correct in the three other reporting categories). However, a detailed analysis of state tests in Texas concluded that such topic reporting is not reliable because problems are not equated for difficulty within these areas (Confrey, Valenzuela, & Ortiz, 2002). Therefore, even though receivers of such reports are being told to be "data-driven" and use the reports to inform their instruction, the MCAS reports themselves are never designed to give feedback at a grain size that could be used for the purpose. A reader can get some intuition on why this is the case by trying problem 19 from the 2003 MCAS shown in Figure 1.2. Then ask yourself "What makes this item difficult?" Clearly, this problem includes elements from four of the 5 "strands" (only missing "Data Analysis, Statistics and Probability"). It is Algebra (for solving equations), Geometry (for its use of congruence), Number Sense (for doing the arithmetic operations), or Measurement (for the use of the

concept of perimeter). Ignoring this obvious overlap, the state chose just one of the 5 strands to classify the problem. It turns out, the state classifies it as Geometry, but later one I will show how our methodology is creating evidence to suggest that there is more to this problem than just Geometry.

		NUMBER OF POSSIBLE POINTS			TOTAL POI (average numl		
					SCHOOL		DIST
		Common	Matrix	Total	#	%	#
REPORTING CATEGORY	Number Sense	13	13	26	11.1	43	10.
	Patterns, Relations, and Algebra	16	16	32	13.0	41	12.
	Geometry	7	6	13	4.[illegible]	38	5.
	Measurement	7	11	18	6.[illegible]	36	5.
	Data Analysis, Statistics and Probability	11	8	19	8.4	44	7.

Figure 1.1. A school level report

19 Triangles *ABC* and *DEF* shown below are congruent.



The perimeter of $\triangle ABC$ is 23 inches. What is the length of side $\overline{DF}$ in $\triangle DEF$?

Figure 1.2. Item 19 from 2003 MCAS test

When we asked this same question about problem 19 to teachers on our Teacher Advisory Board, one of our cooperating educators said "But you know the state - I worked on the MCAS assessment for perimeter [referring to one of

the committees that designs items for the MCAS]... and sometimes when they get a problem that they know, very well, that it could fit more than one strand, they put it in the strand that *they* need to fill. And that is the problem!" Another teacher followed up with "It does affect reports... because then the state sends reports that say that your kids got this problem wrong so they're bad in geometry- and you have no idea, well you don't know what it really is- whether it's algebra, measurement/perimeter, or geometry." A teacher cannot trust that putting more effort on a particular low scoring area will indeed pay off in the next round of testing.

Wiliam (2006) describes an assessment as formative only if information about what is being assessed results in change that would otherwise not occur. Under this guideline, "An assessment of curriculum is formative if it shapes the development of that curriculum. An assessment of a student is formative if it shapes that student's learning" (p. 3). Here the definition of formative assessment becomes more detailed to the utility for teachers to understand if and how students are learning. It would be easier for a teacher to make data-driven changes in her classroom if she had a more detailed analysis of her students' learning. Students' weaknesses need to be addressed by helping them with precise concepts and skills that are neither too easy nor too hard. It was reported that instead of having performance reports that break math knowledge into only a few components, teachers want more fine grained diagnostic reports to accommodate

their everyday classroom practice (e.g., Militello, Sireci, & Schweid, 2008; Wylie, & Ciofalo, 2008). These reports are referred to as "assessment for learning" (Stiggins, 2005; Black et al., 2003).

To help solving these problems, I have been engaged in constructing a fine grained skill model and tracking student's knowledge development at skill level. Meanwhile, I developed an online reporting system to inform teachers about their students' performance on each skill in a timely manner.

1.1.3 Assessment becomes challenging when students learn

Traditionally these two areas of testing (i.e. Psychometrics) and instruction (i.e., math educational research and instructional technology research) have been separate fields of research with their own goals. Standard psychometric theory requires a fixed target for measurement (e.g. van der Linden and Hambleton, 1997), which requires that learning during testing be limited. In psychometrics field, an important assumption of classical test theory (CTT) is that if an examinee could be tested repeatedly using different but strictly parallel forms each time, and assuming test administration was unaffected by previous ones, the average measurement errors would approach zero as the number of administrations approach infinity (Feldt and Brennan, 1993). Item response theory (IRT) (van der Linden and Hambleton, 1997) overcomes some problems of CTT by focusing on

statistical characteristics of individual items. But the local independence assumption still holds (i.e. the response to an item depends only on the item itself and the examinee's ability, but independent of responses to other items) and IRT generally also assumes the examinee ability to be constant and its dimension to be low.

This "fixed target" assumption has hardly met in an ITS as the ultimate goal of a tutoring system is to help students learn. The targets are thereby moving. On the one hand, a tutor often tailors its feedback according to student answers to present question. The feedback is given to students in order to help them learn the content. The tutor also adapts its subsequential activities based on students' answers to previous questions to better remediate problematic knowledge components and, thus, to promote more learning. Therefore, student's knowledge is constantly changing in the intelligent tutor and "local independence" assumption is not valid any more. On the other hand, assessment in an ITS features high dimension. It is common that the number of skills an ITS track in their student modeling module is in the order of hundred instead of just one latent trait. For instance, Cognitive Tutors maintains updated status of student mastery of 2,400 production rules (Ritter, et al., 2009). The knowledge base system GUIDON (Clancey, 1987) represented domain knowledge using about 400 rules and pedagogical knowledge by around 200 rules.

Another characteristic that makes assessment in intelligent tutors special is that different from taking a traditional test, students typically use intelligent tutors regularly in a long term. For instance, in many schools Cognitive Tutors are used 3 days per week over the whole year; similarly, students from Worcester Public Schools would go to school computer lab and work within the ASSISTments system (Razzaq et al., 2005) for 20 minutes every other week as a part of their normal math class. Thus, not only is student ability not fixed during the "assessment" any more, but also it is changing over time. Therefore, the measure of student performance and understanding in such circumstance should be continuous instead of discrete.

Statisticians have done some work to enable assessment of students while they are learning over time. Some attempts to combine standard psychometric models with Markov learning models have been attempted (as far back as Jannarone, 1986). Some work has been done on psychometric measurement of growth and change (e.g. Tan, Imbos and Dos, 1994; Embretson, 1992; Fischer and Seliger, 1997). Gierl and colleagues made contributions in cognitive diagnostic assessment area (Leighton, & Gierl, 2007). Embretson (1991) proposed a multidimensional dynamic latent trait model to account for learning and change. Embretson & Reise (2000) (pp.297-303) provided a comprehensive review on the application of IRT models to study continuous and discontinuous change. However, making predictions from testing data in which students are actively

learning material has only been pursued energetically in the realm of computer-based intelligent tutoring (e.g., Corbett, Anderson and O'Brien, 1995).

One thrust of this dissertation is that I propose approaches to leverage an intelligent tutor to assess students better. I have found evidences showing allowing student learn in the tutor is no longer an obstacle for assessment, but an opportunity to for us to be able to assess more accurately.

1.2 Contribution of this Dissertation

This dissertation makes contributions in several aspects.

(a) First, I propose a novel approach for student assessment. By considering student-system interacting data, I can predict student end-of-year exam score reliably better than traditional methods that pays attention to only correctness (Feng, Heffernan, & Koedinger, 2006a, 2006b, 2009; Feng, Beck, Heffernan, & Koedinger, 2008). My work provides solid evidence for the value of assistance measures, such as the percent correct for scaffolding, the use of help, hint requests, time to respond, and other factors, for prediction of standardized test scores, and suggests that we can do the two things, assessment and assistance, well simultaneously. This work is novel because traditional assessment usually focuses on students' responses to test items and whether they are answered correctly or incorrectly, but ignores all other student behaviors during the test (e.g., response

time). However, an ITS has the potential to use far more. In this dissertation, I take advantage of a computer-based tutoring system to collect extensive and rich data while students interact with the system. The result is shocking: the model (I call it assistance model) that is based upon only the interaction information but includes no correctness on the main problems leads to significantly better prediction of student's state test scores than the model (called lean model) that is based on the correctness alone. The relative success of the assistance model over the lean model highlights the power of the assistance measures. Thus, not only is it possible to get reliable test information while "teaching on the test", data from the teaching process actually improves reliability. This work is novel also in that similar large scale data (learning data of 1000+ students' usage of an online system over two years) has not been available or analyzed beforehand. While the idea of dynamic testing (Brown et al., 1983) is not new, I argue that operationalising it and carefully evaluating whether this promising idea for improving student modeling is achievable in an implemented system that has been widely used by real teachers and students is an important contribution.

The broader impact of this contribution lies in that it has potentially huge practical implications for schools in the U.S. I provide evidence that reliable assessment and instructional assistance can be effectively blended. Thereby teachers can be relieved from the dilemma where they have to make the hard choice between assessment and assistance. While in some sense it is not really

surprising that adding dimensions to the user model improves its accuracy, the underlying debate on assessment vs. teaching makes this data even more relevant. The results with the ASSISTments system open up a the possibility of a completely different approach to assessment that is contentious in nature in suggesting students might not need to spend any time on formal paper and pencil tests.

(b) The second contribution of this dissertation is that I show not only we can hit moving targets, but also we can that longitudinally over a long period of time (Feng, Heffernan, & Koedinger, 2006a, 2006b; Feng, Beck, Heffernan, & Koedinger, 2008). I argue this is novel since no existing systems longitudinally track student knowledge over time and use that to predict student's state test scores. While in this dissertation, I demonstrate how mixed-effects models can be utilized to track change over time and, thus, help do a better job predicting state test scores.

The general implication from this research suggests that continuous assessment systems are possible to build and that they can be quite accurate at helping schools get information on their students over a long period of time. Computer Research Association (CRA) report (Computer Research Association, 2005) has pointed out that continuous assessment systems research is a huge growth area. And, recently, in an interview with U.S. News & World Report

(Ramírez, & Clark, 2009), U.S. Secretary of Education Arne Duncan also called for continuous assessment. He mentioned that he is concerned about over-testing, and feels that fewer, better tests would be more effective. Many states are moving towards adopting "*value added*" assessments, so that they can track the *value added* by teachers and schools. *Value added* is possible because you have year to year state assessments so you can see the average learning gain for students per year, and attribute those gains to teachers and schools. Such systems could benefit from data that is collected every few weeks, instead of once a year, thereby allowing schools to more quickly figure out *what works* at increasing student learning and get more accurate estimates of student performance later as well.

(c) This dissertation makes a unique contribution by developing skill models of various granularities and rigorously evaluates the effectiveness of the models at estimating student knowledge (Feng, Heffernan, Mani, & Heffernan, 2006; Feng, Heffernan, Heffernan, & Mani, 2009). One of the main contributions of this dissertation is that it demonstrates the value of a very fine-grained versus more coarse-grained models within intelligent tutoring systems. I rigorously evaluate the effect of granularity of the skill models and validated the usage of a fine-grained skill models. While some prior research in the field of intelligent tutoring systems has looked at the value of coming up with accurate skills models (and has generated some automated methods for doing so, such as Barnes, 2005; Cen, Koedinger & Junker, 2005, 2006), my work is different in that we hand-coded the

skill models and built the connection between skills and questions. This is similar to what Ferguson et al. (2006) did as they also associated problems with skills by hand, but they employed a different methodology. Another contribution of this dissertation is that I demonstrate how skill models in intelligent tutoring systems can be used to predict standardized test scores.

The practical implication of this contribution has to do with the balance between informative feedback and accurate assessment. Traditionally, most assessments, including the high stake state tests required by NCLB, treat all questions on the test as sampling a single underlying knowledge component. Yet, the fact is that instead of having performance reports that break math knowledge into only a few components, teachers want more fine grained diagnostic reports to accommodate their everyday classroom practice (Militello, Sireci, & Schweid, 2008; Wylie, & Ciofalo, 2008). Can we have our cake and eat it, too? That is, can we have a good overall prediction of a high stakes test, while at the same time be able to tell teachers meaningful information about fine-grained knowledge components? The answer from the dissertation is "yes". I show that we can inform teachers of fine grained skills with which their students are having trouble as well do a better job of predicting student's knowledge using the fine grained model than using coarser grained models, which is surprising because for each skill we have less information.

(d) I implement a reporting system that has been deployed and used by real teachers (Feng & Heffernan, 2005, 2006, 2007), as part of the ASSISTment system. The complete working system, ASSISTments system, is available on the World Wide Web (www.assistment.org) and has been used by more than 6,000 students (Razzaq, Feng, et al., 2005, 2007). The system makes it possible to assess students and assist student learn at the same time. This reporting system is different from reports from state tests in that first, the reports are made available online lively. Thus, instead of waiting for six months, teachers can view the reports while students working on ASSISTments; secondly, the system reports to teachers based upon a very fine grained skill model as opposed to the state test reports that only break math into only a few knowledge components, so that teachers know what's really causing difficulties to their students.

The U.S. Department of Education (2003) stated, "Research shows that teachers who use student test performance to guide and improve teaching are more effective than teachers who do not use such information" (p. 2). Therefore, if assessments are to move from assessments *of* learning to assessments *for* learning (c.f., Stiggins. 2005) then we must continue to focus on the box that encompasses diagnostic modeling data and teachers as the end-users. By giving teachers cognitively diagnostic data in a timely fashion through my reporting system, teachers can change their teaching in response to the data they collect about student understanding.

(e) I establish an easy approach to estimate and compare the effectiveness of different tutoring content (Feng, Heffernan, & Beck, 2009). The traditional approach uses randomized controlled studies that can be very expensive speaking of the time and subjects, especially if we want to focus on hundreds of individual pieces of tutoring contents. Yet, I demonstrate an approach where we can apply educational data mining technique to use observational data to efficiently compare the effectiveness of individual pieces of instructional materials in a tutoring system. This is a low cost approach comparing to randomized experiments. My analyses show in the ASSISTments system, some contents have more impact on student math skill development while some contents are not so useful. I suspect this result is not specific to ASSISTments, and other tutors may have items that vary greatly in educational effectiveness. Potentially, the method can be used generally to examine the quality of instructional contents in learning systems, and thus improve the overall learning impact.

To summarize, as an interdisciplinary research, this dissertation contributes to several academic domains. First, the dissertation contributes to the field of computer science, particularly**, intelligent tutoring systems and user modeling**. Chapter 3 contributes to the field of intelligent tutoring systems by aiming to exploit the fine grained data in a tutoring system to improve the accuracy of modeling a learner. In particular, where many uses of assessment use only the overall correctness of the student's answers, a tutoring system has the potential to

use far more. The results indicate that not only can a tutoring system be used to help student learn which has been the goal of most systems, but also it can be quite accurate at assessing students. Following approaches provided by this dissertation, intelligent tutoring system builders can build detailed learner models to "feed" into the tutoring process that can be used to inform teachers, students and the machine on the ways to best remedy problems. This part of the dissertation was published and well-received at World Wide Web Conference (WWW) (nominated for best student paper award) and the journal of User Modeling and User Adaption and Interaction (UMUAI) (nominated for James Chen award, best UMUAI paper of the year).

The topic of user modeling is important, especially with the increasing role of standardized testing. In general, being better able to diagnose a student's knowledge is critical for improved and targeted teaching, for both teachers and intelligent tutoring systems. Chapter 4 makes significant contribution on this aspect. Early in 1994 when speculating on long term research goals, researchers in intelligent system field, McCalla & Greer (1994) pointed out that the ability to represent and reason about knowledge at various levels of detail is important for robust tutoring. As far as I know, my contribution is entirely unique in that it rigorously evaluates the effect of the level of details within the skill models, using usage data of about 1,000 students from over two years from a tutoring system and external state test scores, going from very broad to very detailed. Since

intelligent tutoring systems tend to use fine-grained skill models, the work thus validates a core underlying assumption. This part of the dissertation was published in IEEE Transaction on Learning Technologies (featured article of the issue).

Researchers in **data mining** area has been searching for automated techniques to address interesting scientific questions, such as revising transfer models for better knowledge representation. Yet, the work in Chapter 5 starts with human constructed models and ends with the conclusion that students of different proficiency level can be best represented by different transfer models. I was also trying to bring together efforts from human subject experts, educational data mining approach, and intelligent searching process from artificial intelligence field to refine existing transfer models. I think this is an important piece work in that it helps us with better understanding what parts of the scientific enterprise can be best done by people and which are better done computationally, which is a major issue in the area of educational data mining.

This dissertation makes important contributions to educational measurement, esp. **psychometrics and educational assessment development**. Assessment models in the psychometrics society usually assume a fixed target during a test in a limited time, which is always violated in a tutoring system. How to hit a moving target thereby becomes an important issue to be addressed because we want

students to “move” (i.e. they learn and their ability improves). Yet, I establish an approach showing because our system not only teaches while it assesses, but it even does a better job of assessing because of learning, and longitudinal modeling approach allows us to do so over a long period of time. As nowadays more and more computer-based tests are being administered in replacement of paper-and-pencil tests, the results from this dissertation could well influence the development of e-assessment (or adaptive assessment) system. They bring up a good reason for psychometricians to develop models that pay more attention to the data that they have discarded in the past, such as students’ response time, their attempting frequency, their learning over time, and so on.

1.3 Dissertation Organization

The remainder of this dissertation is organized into 6 chapters. The second chapter gives an overview of the ASSISTments system. In particular, it shows how students work in the ASSISTments.

The third chapter focuses on the reporting system in the ASSISTments project. It describes representative reports that have been built to inform teachers and shows how one of the reports promotes later works of the dissertation. In addition, teachers’ attitude towards the system will be discussed.

The forth chapter addresses the testing challenge in ASSISTments. Particularly, in this chapter, I present the dynamic metrics that I develop to measure the amount of assistance students need to solve problems. After analyzing usage data of 1, 000 students in two years, I argue that the dynamic metrics are more helpful at predicting students' state test scores than the correctness of their answers. I also show how to longitudinally model their knowledge development in an environment where students learn. Additionally, methods are proposed to evaluate our predictions. I argue that the accuracy of the predictions is comparable with the standardized test itself.

The fifth chapter reports my work on developing and evaluating fine grained skill model. In the chapter, I focus on explore two research questions: Does adding scaffolding responses help? And does the finer-grained model predict better than coarser grained models? The modeling process will be described on how I evaluated models of various granularities based on data of more than 1,000 students and their external state test scores. Solid evidence is presented that the fine grained model leads to more accurate estimates of students' state test scores.

Chapter six discusses my effort on improving the system, including detecting instructional differences among individual contents in tutoring efficacy and refining existing skill models based on findings from data analysis. Conclusions of this dissertation and the future work are described in Chapter seven.

Chapter 2

2 Background

2.1 Overview of the ASSISTments System

Back in 2003, the US Dept of Education funded Heffernan and colleagues to build a system that assesses 8th grade math while it tutors students. Since spring, 2004, I, together with other colleagues, have been working on the ASSISTments system that was built to offer instruction to students while providing a more detailed evaluation of their abilities to the teacher than is possible under current approaches. The key feature of ASSISTments is that they provide instructional assistance in the process of assessing students. If students got the item correct they were given a new one. If they got it wrong, they were provided with a small "tutoring" session where they were forced to answer a few questions[4] that broke

[4] As future work, once we have built a predictive model and are able to reliably detect students trying to "game the system", (i.e., just clicking on answer) we will allow students to re-try a

the problem down into steps to eventually get the problem correct. During the tutoring session, students are allowed to request for hint messages and make more than one attempts at the questions.

The intelligent tutor was deployed with a completely internet savvy solution and developed based on the Common Tutoring Object Platform (CTOP) (Nuzzo-Jones, Walonoski, Heffernan & Livak, 2005) whereby students access the web site via a web browser that reported each student's actions on the web page to our database at WPI, thus enabling completely live database reporting to teachers, as I will describe in Chapter 4. The ASSISTments system is about 5 years old and has been primarily used by middle-school and high-school teachers and students throughout Massachusetts who are preparing for the MCAS tests. The system started by interviewing a few middle school math teachers about how they would tutor a problem. In May of 2004, one week before the MCAS was given, about 300 students participated when 5 different teachers brought their students to their schools' computer labs. In the 2004-2005 school year, more than 600 students used the system regularly about every two weeks. Eight math teachers from two middle schools would bring their students to the computer lab, at which time students would be presented with randomly selected MCAS test items. Currently,

question if they don't seem to be "gaming" (Baker, Corbett, & Koedinger, 2004). Thus, studious students will be given more flexibility.

there are over 3,000 students and 50 teachers using our system as part of their regular math classes. The ASSISTments is a collaborating project between researchers and teachers: now over 30 teachers have used the system to create contents; 17 teachers are participating a workshop at WPI to learn how to make more effective use of diagnostic assessment data from ASSISTments; Heffernan and colleagues, and WPI graduate students still visit schools regularly to help with technical issues and on how to use ASSISTment data to drive classroom instructions.

The individual pieces of tutoring for a given MCAS question are called "ASSISTments", a word coined by Kenneth Koedinger, because they blend *assist*ing the student to learn with assess*ment*. Each ASSISTment consists of an original question and a list of scaffolding questions. The original question usually has the same text as in MCAS test while the scaffolding questions were created by our content experts to coach students who fail to answer the original question. The first scaffolding question appears only if the student gets the item wrong. An ASSISTment that was built for item 19 of the 2003 MCAS is shown in Figure 2.1. Particularly, Figure 2.1 shows the state of the interface when the student was partly done with the problem. We see that the student typed "23" (which happened to be the most common wrong answer for this item from the data collected). After an error, the textbox was then disabled and the student was not allowed to try the item further, but instead must then answer a sequence of

scaffolding questions (or "scaffolds") presented one at a time. Students work through the scaffolding questions, possibly with hints, until they eventually get the problem correct. If the student presses the hint button while on the first scaffold, the first hint is displayed, which would be the definition of congruence in this example. If the student hits the hint button again, the second hint appears which describes how to apply congruence to this problem. If the student asks for another hint, the answer is given. Once the student gets the first scaffolding question correct (by typing "AC"), the second scaffolding question appears. Buggy messages will show up if the student types in a wrong answer as expected by the author. shows a buggy messages that appeared after the student clicked on "½*x(2x)" suggesting he might be thinking about area. (Hints appear on demand, while buggy messages are responses to a particular student error and will show up on the screen automatically) Once the student gets this question correct he will be asked to solve 2x+x+8=23 for 5, which is a scaffolding question that is focused on equation-solving. So if a student got the original question wrong, what skills should be blamed? This example is meant to show that the ASSISTments system has a better chance of showing the utility of fine-grained skill modeling due to the fact that we can ask scaffolding questions that will be able to tell if the student got the question wrong because they did not know congruence versus not knowing perimeter, versus not being able to set up and solve the equation. As a matter of

logging, a student is only marked as getting the item correct if he/she answered the questions correctly before asking for any hints or encountering scaffolding.

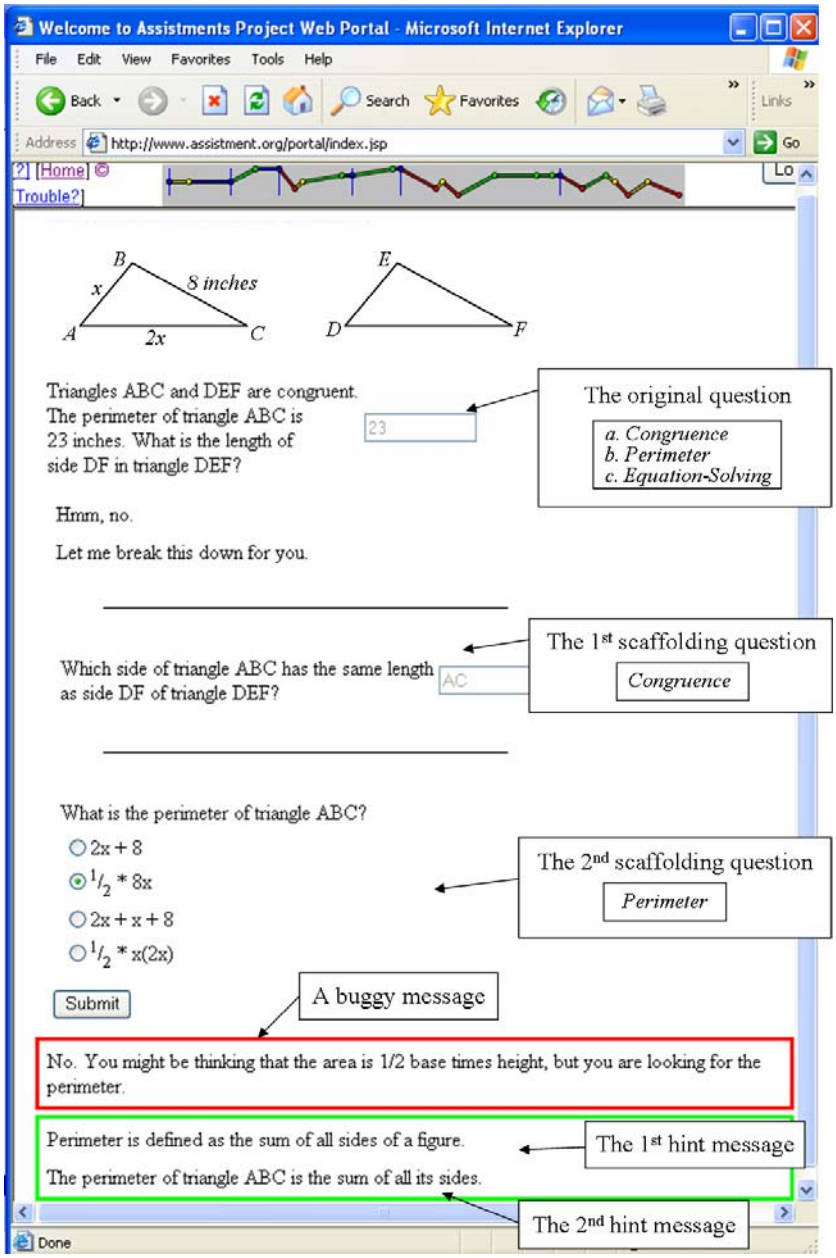


Figure 2.1. An ASSISTment shown after a student hits the "Hint" button twice, showing the two different hints and one buggy message that can occur at different points

A key hypothesis of this project is that ASSISTments can do a better job of assessing student knowledge limitations than practice tests or other on-line testing

approaches by using a "dynamic assessment" approach (Grigorenko & Sternberg, 1998); in particular, use the amount and nature of the assistance that students receive which is normally not available in traditional practice test situations as a way to judge the extent of student knowledge limitations. These measures include the time it takes students to come up with answer to a question, the attempts they take to correct an answer if it is wrong, the number of hints they request, and their performance during the tutoring session on the scaffolding questions. My effort to test this hypothesis of improved prediction of the ASSISTments' dynamic assessment approach is discussed in Chapter 4. Further, because the scaffolding questions break the test questions down into a series of simpler tasks that directly assess fewer knowledge components, I hypothesize that I can do a more accurate assessing job. This hierarchal breakdown will provide a much finer-grained analysis than is currently available for any large-scale tests, such as MCAS. Moreover, since students use the web site regularly during a year, we can track student performance development longitudinally. Each week, the web site will learn more about each student and will provide an increasingly accurate prediction of how well each student will do on the MCAS.

ASSISTments is definitely a team effort. My responsibility in this project includes the development of the reporting system, and later on analysis of data collected from the system. It is worth mentioning that all the data used in this dissertation have been collected from the ASSISTments system. It is the special

features of ASSISTments that make it possible for us to do a better job assessing students.

2.2 The Massachusetts Comprehensive Assessment System

The Massachusetts Comprehensive Assessment System (MCAS) is a high-stakes testing system required by the *No Child Left Behind Act*. In Massachusetts, MCAS is the graduation requirement in which all students in state educated with public funds in the tested grades are required to participate. It is administered as standardized test that produces rigorous tests in English, math, science and social studies for grades 3 to 10 every year. Students need to pass the math and English portions of the 10^{th} grade versions in order to get a high school diploma. The test is filled with challenging multi-step problems that tap a variety of different mathematical concepts. Students even have to write essays explaining the mathematics they use to solve problems. (See Appendix B for sample released MCAS test items from the year 2003).

MCAS results for students are reported as scaled scores and performance levels that are calculated based on the raw scores. The raw score reflects the total number of points a student got from a MCAS test, ranging from 0 to 54. The range of scaled scores for MCAS tests is arbitrarily defined to be between a

minimum of 200 and maximum of 280, which is further divided into four performance levels, namely Warning/Failing (when the scaled score falls in the range of 200 ~ 218), Needs Improvement (220 ~ 238), Proficient (240 ~ 258), and Advanced (260 ~ 280). The purpose of scaled scores is to provide information about the position of a student's results within a performance level. Student raw scores, or the total number of points, on the MCAS tests are converted into scaled scores using a scaling procedure. Generally speaking, the MCAS scaling procedure involves two-step transformations: "*(1) non-linear monotonic transformations of the raw score points into theta metric*[5] *points, and (2) linear transformations of theta points into MCAS scaled score points.*" (MCAS technical report, 2001, p41) These transformations do not change the rank ordering of students, nor their performance level classifications. Also, three threshold raw scores determined by a standard setting procedure were used as the cut points for the four performance levels. The whole procedure tries to minimize the fluctuations in the thresholds from year to year, consistent with the slight shifts in the difficulty of the tests and the gaps in theta estimates between rounded raw scores.

My dissertation work relates to the MCAS in several ways. First, MCAS helps set the goal for the ASSISTments project. Because students are more likely to fail

[5] The "theta metric" refers to the student ability (θ) metric in Item Response Theory models.

the mathematics portion of the test[6], the state is focusing efforts on mathematics. 8th grade math thereby became where the ASSISTments project started to help students get better prepared for the MCAS[7]. Secondly, we have built our content based upon the released items from previous MCAS math tests. In Massachusetts, the state department of education has released publicly 11 years worth of 8th grade MCAS test items on math, totalling over 400 items, which we have turned into ASSISTments by adding "tutoring". The items that have been used include not only the publicly available released items but also variations on these problems involving different numbers, cover stories, or combinations of concepts and skills. By using MCAS released items we both help to encourage transfer to the real test, as well appear to parents, teachers, and students to be a worthwhile activity that is geared towards the MCAS objectives that they care about. Thirdly, student MCAS scores have been widely used in the student response data modelling process of my research. For instance, I evaluate models by examining how well they predict the MCAS test scores which was taken by students at the end of a school year after the online data being used was collected, as I will describe later in Chapter 4. Fourthly, the Massachusetts Curriculum Framework from MCAS provides a base for the

[6] http://www.edweek.org/ew/newstory.cfm?slug=02mcas.h21

[7] We started with 8th grade mathematics because it is the "hard" part. The system itself was built in the way that it is not limited to a specific domain or a particular grade. Now the system also includes 6th grade and 10th grade math contents. Meanwhile it has expanded to tutor science inquiry skills (Sao Pedro et al., 2009).

construction of our fine grained skill model (See Appendix C for learning standards for Grades 7-8). It also serves as contrast cases for my analysis of effectiveness of skills model at various granularities, as I will describe in Chapter 5.

Chapter 3

3 Informing Teachers Live

Chapter organization: In this chapter, section 1 discusses more about the motivation of building the reporting system. Section 2 presents representative reports that I have designed and implemented to give teachers feedback on their students' performance and to help them analyze the contents. Section 3 shows teachers' attitude towards the reporting system, and in Section 4, reports from other systems are described and contribution of my reporting system is discussed. The content of this chapter draws heavily from the paper (Feng, & Heffernan, 2005) that has been presented at a workshop, and two journal papers (Feng, Heffernan, 2006, 2007).

3.1 Introduction

As discussed in Chapter 1, schools seek to use the yearly MCAS assessments in a data-driven manner to provide regular and ongoing feedback to teachers and students on progress towards instructional objectives. Plus teachers and parents

also want better and faster feedback than they currently receive. Among other things, the feedback on the MCAS is broken down into only 5 strands despite the fact that the number of mathematics skills and concepts that a student needs to acquire is on the order of hundreds. In Chapter 2, I have shown an item was filled into Geometry strand though it could naturally fit more than 1 strand.

The question of tagging items to learning standards is very important because teachers, principals and superintendents are all being told to be "data-driven" and use the MCAS reports and use the MCAS reports such as in Figure 1.1 to adjust their instruction. However, these MCAS reports were never designed to give feedback at a grain size that could be used to guide instruction. The MCAS, and almost all other similar state tests (as well as many tests like the GRE and the SAT) are as well, unidimensional tests that assume there is only one "latent" (i.e. unobserved) skill that is being tapped (i.e. the one skill of "MATH" as opposed to tracking algebra and geometry separately). Such tests work pretty well at telling you which students are performing well since students that do well in one area tend to do well in others, but are not good at *informing educators* about how to help students. There are several reasons for this poor MCAS reporting: 1) the reasonable desire to tag problems with multiple skills, 2) the fact that paper and pencil tests cannot figure out, given a student's response, which skills to credit or blame, 3) there are skills that deal with decomposing and recomposing multi-step problems, yet are currently poorly understood by cognitive science. It would be

easier for a teacher to make data-driven changes in her classroom if she had a more detailed analysis of her students' learning on a finer grained knowledge level and the analysis results are reported periodically, or better lively, along the school year.

To address the reporting issue with MCAS test, the primary goal of this reporting system is to assist student learning by providing feedback of precise assessment to teachers, so that they can adjust their classroom strategies based upon the system's fine-grained analysis of student learning. I conjecture that detailed reporting with precise assessment, closely tied to the material students are working on, will make it easier for teachers to use data-driven decision makers to alter their planned instruction in response to the need of the majority of the class and to the need of individual student. The reporting in the ASSISTment system was built to identify the difficulties individual students – and the class as a whole – are having. It is intended that teachers will be able to use the detailed feedback to tailor their instruction to focus on the particular difficulties identified by the system. Compared to the MCAS reports, reports provided by the reporting in the ASSISTment system is live so that teachers do not need to wait. We have built different cognitive models and reported to teachers in term of finer grained-sized skills in addition to the learning standards. The cognitive model we built allow one problem to be tagged with multiple skills and finer grained models that break down the 5 strands into about 100 skills and code the problems (also the

scaffolding questions) with the new skills. I will describe the fine grained cognitive model in further detail in Chapter 5 where the creation, analysis and refinement of cognitive models will be discussed.

3.2 The ASSISTments Reporting System

The ASSISTment system was deployed with a completely internet savvy solution that reported each student's actions to our database at WPI as they are working in the system, thus enabling completely live database reporting to teachers (Feng & Heffernan, 2005, 2006, 2007b). Teachers seem to think highly of the ASSISTment system not only because their students can get instructional assistance in the form of scaffolding questions and hint messages while working on real MCAS items, but also because they can get online, live reports on students' progress while students are using the system in the classroom. Below in this Chapter, I present several representative reports as a part of the reporting system that I built to help inform teachers.

3.2.1 Real-time lab performance monitoring

The "Grade Book", shown in Figure 3.1, is the most frequently used report by teachers. Each row in the report represents information for one student, including how many minutes the student has worked on the ASSISTments, how many minutes he has worked on the ASSISTments today, how many problems he

has done and his percent correct, our prediction of his MCAS score and his performance level. Our "prediction" of a student MCAS score is at this point primitive. The column is currently simply a function of percent correct. Back in 2005, we found a strong correlation between our prediction for the 68 students who have used our system May 2004 (the first group of users of ASSISTment) and their real MCAS raw score[8] (r = 0.7) (Razzaq et al., 2005). And I was continually refining our prediction function based on new data (Feng, Heffernan, & Koedinger, 2006a, 2006b, 2009). My work on developing better predictive models for MCAS will be discussed in Chapter 4 and Chapter 5.

[8] MCAS test results are reported in the form of scaled scores and performance levels for individual students, schools, and districts. Individual student raw scores are translated into scaled scores and performance levels through a process called scaling. The raw score, ranging from 0 to 54 will be converted into scaled score which ranges from 200 to 280. The total range of MCAS scaled scores is divided into four equal segments, each referring to one performance category, namely warning/failings, need improvement, proficient, and advanced. Please refer to http://www.doe.mass.edu/mcas/ for further details.

Student Name	Elapsed time (hh:mm)	Original Items				Scaffolding + Original Items		
		# Done	% Correct	Est. MCAS Scaled Score*	Est.MCAS Performance Level	# Done	% Correct	# Hint Req.
Tom*	4:12	90	39%	214	Warning/Failing	228	44%	353
Dick*	4:01	98	66%	244	Proficient	158	59%	58
Harry*	4:07	78	40%	224	Needs improvement	154	38%	77
Mary*	4:17	114	20%	200	Warning/Failing	356	20%	705
Jack*	3:53	104	39%	214	Warning/Failing	267	43%	187
John*	4:24	92	55%	244	Proficient	40	52%	55

Figure 3.1. A grade book report on real student data for one of our teachers' classes

In Figure 3.1, we see that these 6 students have used the system for about 4 hours. "Tom" has finished 90 original items. Many of those he got wrong and so was asked many scaffolding questions of which he got 44% of the 228 questions asked correct. He asked for 353 hints. Our prediction of his MCAS score was relatively low, "Warning/Failing". Teachers can also see "Mary" has asked for way more hints than others (705 hints). Noticing this, a teacher could go and confront the student with evidence of gaming or give him a pep-talk. The "Grade Book" is so detailed that a student commented: "It's spooky", "He's watching everything we do" when her teacher brought students to his workstation to review their progress. By clicking the link next to each student, the teacher can see what those questions were and what kind of errors the student made. Knowing students' reactions to questions helps teachers to improve their instruction and enable them to correct students' misunderstandings in a straightforward way.

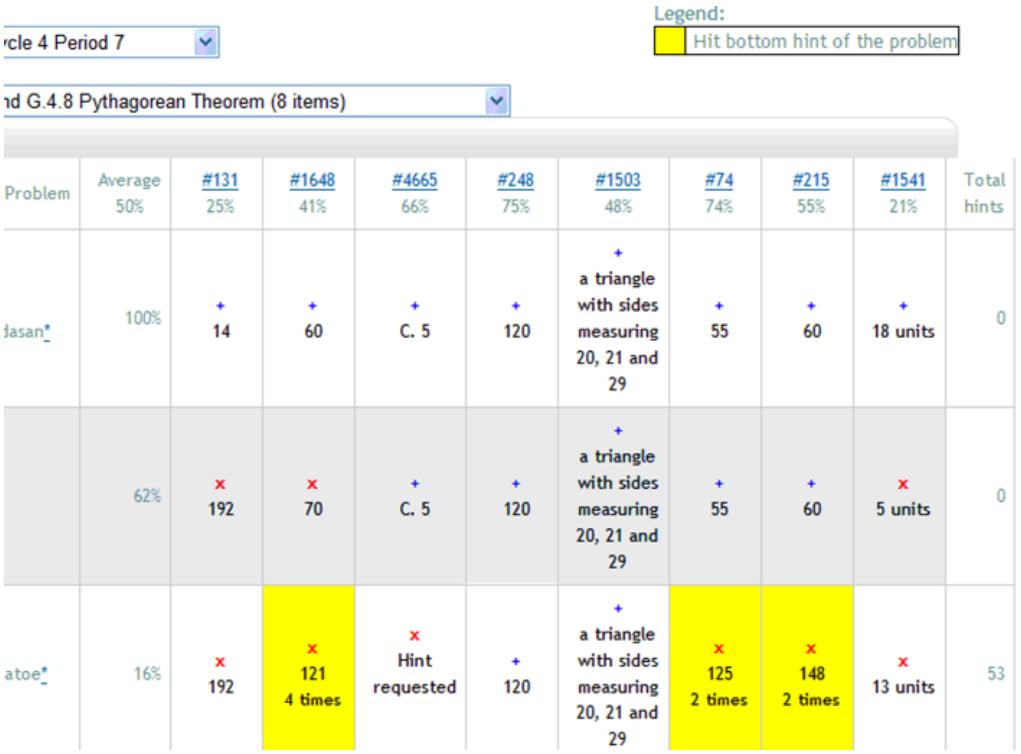


Figure 3.2. An item report tells teachers how students are doing on individual problems.

The grade book report gives an overview of a student/a class's performance. Figure 3.2 shows an item report which is generated based on detailed, action-level logs of students. The report shows teachers how students are doing on individual problems. In particular, the item report in Figure 3.2 presents how three students from one class did on an assignment of eight problems focusing on Pythagorean Theorem, a skill that the students are requested to acquire in the eighth grade. By presenting in different colours and using different tags, the report helps teachers quickly tell if a student answered the question correctly (indicated by a "+" sign in

the report); if not, did they give incorrect answer (indicated by an "x" sign) at their first attempt or they requested for hint (indicated by a message "Hint requested"). For instance, the three yellow-highlighted cells in the third row leap out at the first sight. Hence, the teacher can tell at a glance that the student is asking for too many bottom-out hints. Actually, the student requested for hint messages 53 times, far more than others. Teachers can also see what students have answered for each question. As shown in the report, the correct answer of item #131 is "14", yet two of the three students gave the answer "192", which could be a warning signal to the teacher suggesting there are some common misunderstandings among students. Through providing the detailed information, we hope to help teachers inform their instructions and, thus, provide better remediation to students.

Teachers seem to think highly of the ASSISTment system not only because their students can get instructional assistance in the form of scaffolding questions and hint messages while working on real MCAS items, but also because they can get online, live reports on students' progress while students are using the system in the classroom.

3.2.2 Help identifying difficult steps

Teachers want to know particular skills or knowledge components that cause trouble to students. Unfortunately the MCAS is not designed to be cognitively

diagnostic. By breaking original items into scaffolding questions and tagging scaffolding questions with knowledge components, we are able to analyze individual steps of a problem. Figure 3.3 is what we call a scaffolding report because it reports statistics on each of the scaffolding questions that are associated with a particular original item.

ID	Question	Correct Answer	% Correct	Hint Req.	# Attempt	Common Errors			WPI's Use of MA. Standard	WPI's Knowledge Components
						Resp.	#	Buggy Message		
Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches. What is the length of side DF in triangle DEF?		10	12%	56%	180	8	15	N/A	**G.2.8**, M.3.8, P.7.8	Composition, T.3, A.3, T.4, A.4, A.12, A.15, A.17
						16	13	N/A		
						23	8	N/A		
1	Which side of triangle ABC has the same length as side DF of triangle DEF?	ac	23%	50%	154	ab	13	Side AB corresponds to side DE of triangle DEF, not DF. Try again, please.	**G.2.8-congruence-and-similarity**	Term: "Congruency", Appl: Congruency
						DF	6	N/A		
2	What is the perimeter of triangle ABC?	2x+x+8	39%	20%	148	2x + 8	69	No. It looks like you have added just two of the sides of triangle ABC. Perimeter is the sum of all the sides.	M.3.8-using-measurement-formulas	Term: "Perimeter", Appl: Perimeter
3	Now, given the perimeter of triangle ABC equals 23 inches, you can write the equation **2x + x + 8 = 23** and solve it for **x**. What is the value of **x**:	5	25%	52%	147	15	13	N/A	P.7.8-setting-up-and-solving-equations	Appl: Solve linear equation
						13	10	N/A		
						8	10	N/A		
4	Remember, we are looking for side DF. Enter the length of side DF:	10	30%	43%	143	5	26	N/A	G.2.8-congruence-and-similarity	Appl: Congruency
						2x	2	N/A		
						8	3	N/A		

Figure 3.3. A "scaffolding report" that teachers can generate already from the reporting system

On the first line of Figure 3.3, we see this problem is hard since only 12% of the students got it correct on their first attempt. Of the 180 students having done this item so far, 88% (about 158 students) could not get the correct answer to the original question, thus forced by the system to go through scaffolding questions to eventually solve the problem. Eagle-eye readers may notice that 154 (the number of attempt on the first scaffolding question) is less than 88% of 180. And the

number of attempts on later scaffolding questions went down more. That's because students could log out and log back in to redo the original question to avoid going through all scaffolding questions. This problem has been solved. 56% of students asked for a hint, telling you something about students' confidence when confronted with this item. (It is useful to compare such numbers across problems to learn which items students think they need help on but don't, and vice versa). Remember that the state classified the item into "Geometry" category according to its "congruence" (G.2.8) shown in bold. The other MA learning standards (M.3.8, P.7.8) are the learning standards we added in our first attempt to code using the MCAS 39 standards. We see that only 23% of students that got the original item incorrect can correctly answer the first scaffolding question lending support to the idea that congruence is tough. But we see a as low percent correct 25% on the 3rd question that asks students to solve for x. The statistics result gives us a good reason to tag "P.7.8-setting-up-and-solving-equations" which is a learning standard under the strand of "Patterns, Relations, and Algebra" to the problem.

Given the scaffolding report can provide lower level of cognitive diagnosis, our cooperating teachers have carefully designed scaffolding questions for those tough problems to find out the answer. For example, one teacher designed an ASSISTment for ("What's ¾ of 1 ½?"), item 20 of year 2003 8th grade MCAS. The first scaffolding question for the ASSISTment is "what mathematical

operation does the word ‘of’ represent in the problem”. This teacher said, “Want to see an item that 97% of my students got wrong? Here it is… and it is because they don’t know ‘of’ means they should multiply.” The report has confirmed the hypothesis. 40% of students could not select “multiplication” with 11 of them selecting “division”.

The scaffolding report has helped us to develop our tutors in an iterative way. For each question, the report shows top common errors and corresponding “buggy” messages. When building the scaffoldings for ASSISTments, we have tried to “catch” common errors students could make and give them instructive directions based on that specific error, such as correcting students’ misunderstanding of question texts or knowledge concepts. But given that students may have different understandings of concepts, ASSISTments may give no messages for some errors, which means our tutor lost chances to tutor students. Also, students may feel frustrated if they are continually being told “You are wrong” but get nothing instructive or encouraging. As shown in Figure 8, the wrong answer “15” to the third question has been given 13 times, but the ASSISTment gave no instructive messages. Noticing this, the ASSISTment builders can improve their tutor online by adding a proper “buggy” message for this error.

Original	154																22
Q1	119								35								N/A
Q2	85				34				12				23				
Q3	72		13		21		13		8		4		18		5		
Q4	45	8	5	7	15	6	3	10	6	2	1	3	15	3	1	4	

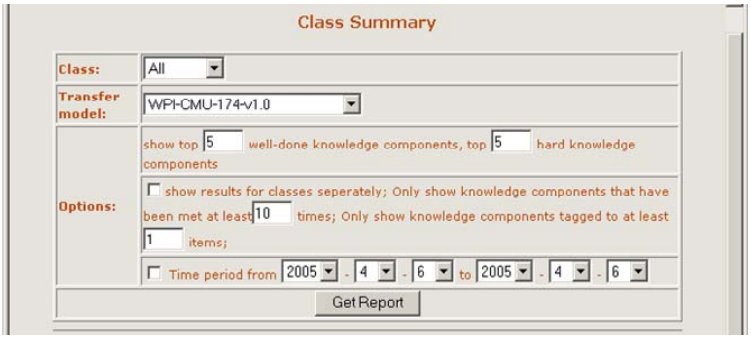
Figure 3.4. Red & Green" distribution matrix

I also display a table that I call "Red & Green" distribution matrix as shown in Figure 3.4 following the scaffolding report. Numbers in the cells show how many students got correct (indicted by green number in un-shaded cells) or wrong (indicated by red in shaded cells) on a question. The number is split as the questions' sequence number grows so that it also represents how those students have done on previous questions. In this example, we see that 4 students who have answered the original question wrong went through all of the scaffolding questions correctly. Given that, we tend to believe those students have mastered the knowledge components required by each step and but need instruction on how to "compose" those steps. It's also worth pointing out that there are 8 students who answered original question wrong but answered correctly to the last question, which asks the same question as the original one. Since the assistment breaks the whole problem into scaffolding steps and gives hints and "buggy" messages, we would like to believe those students learned from working on the previous steps of this assistment.

3.2.3 Informing hard skills

Tagging questions with knowledge components in different transfer models enables us to develop reports to inform teachers about the knowledge status of

classes and of individual student. “Class Summary” report and “Student-Level Knowledge Component” report were developed for this purpose. As shown in Figure 3.5, teachers can select their favourite transfer model, specify the number of knowledge components to be shown in the report. Knowledge components are ranked according to their correct rate which is students’ correct rate (demonstrated in Figure 3.5 as green bars together with percent correct as values) at the items tagged with those knowledge components. By clicking the name of a knowledge component (shown as a hyperlink in Figure 3.5), teachers are redirected to another page showing the items tagged with the knowledge components. In the new page, teachers are able to see the question text of each item and continue to preview or analyze the item if they want to know more about the item.



5 skills your students doing well

WPI-5	WPI-39	WPI-78	Correct Rate		
Number Sense	N.10.8-computing-numbers	Addition		86%	392
	N.1.8-number-representations	Integers		85%	107
		Ordering-Numbers		79%	312
		Rounding		79%	164
	N.10.8-computing-numbers	Subtraction		76%	715

5 skills your students need more practice

WPI-5	WPI-39	WPI-78	Correct Rate		
Measurement	M.2.8-concerting-measures	Unit-Conversion		44%	1063
Algebra	A.7.8-solving-equations	Inequality-Solving		43%	593
Geometry	G.1.8-polygon-geometry	Sum-Of-Interior-Angles		42%	124
	G.4.8-pethagorean-theorem	Pythagorean-Theorem		37%	890
Data Analysis	D.2.8-data-representation-tech.	Venn-Diagram		37%	424

Figure 3.5. The class summary report (settings and results)

By presenting such a report, we hope we can help teachers to decide which knowledge components and items should be focused on to maximize the gain of students' scores at a class level when instructional time is limited. We would like to evaluate the effectiveness of the report by comparing the learning gain in a limited time of the classes for which the teachers have been exposed to this report to the control groups for which this report is not accessible.

In addition to the class level knowledge component report, we present a student level report (developed by Quyen Do Nguyen at WPI) to teachers which

shows the knowledge status of individual students. Similar to the class level report, strong and weak knowledge components are listed, but only for the particular student specified by the teacher. The student level knowledge component report comes after the class level report and quickly becomes a favourite report of our cooperating teachers. Teachers love the fact that they can see in this report the weak points of a particular student in their classes so that they will pay more attention to those knowledge components when giving instructions to the student. Since both original items and scaffolding steps have been tagged in different grain sized transfer models in the ASSISTments system, I argue that we can more accurately detect what are the real obstacle knowledge components for each student.

3.3 Teachers' Attitude

Nice things have been mentioned about the system by our cooperating teachers anecdotally. To collect usage feedback from teachers, we created an online survey of teacher attitude about the ASSISTment system and how do they use the data from the system during the school year 2005 to 2006. The responses were positive. Teachers in general liked the feature that the ASSISTments lead students step by step when they incorrectly answered a question and "it was great to have the 'hints' that are tailored to their individual needs". They also considered using the system as a good MCAS practice and loved the fact that they

can receive scores at the end of the class. Among the 11 teachers who responded, 8 teachers strongly agreed that they thought their students learned by using the system and 3 agreed somewhat. And nine of the teachers would consider assigning ASSISTment problems as homework for students with computers at home.

I have noticed a discrepancy that although eight of the 11 teachers thought the data provided the system was helpful, only 3 teachers said that they did use the data (i.e. the reports) to change what they did instruction in class while 7 others mentioned that they only did this somewhat. I hypothesize that one reason for this difference can be the availability of the reports. Most teachers are not accustomed to frequently log in to the system to access the reports on their own. To some teachers, doing this also adds extra work. Actually when being asked the opinion on receiving automatic email reports, 9 teachers thought that would be great since it would be "a much easier and faster way of obtaining the information" and it would eliminate work for them so that allow them "more time to focus on certain strategies or concepts in class".

Another thing we care about is how the teachers use the reports. In the survey, most teachers only mentioned that they reviewed common mistake problems with the whole class, which indicated that many functions provided the reporting system have been ignored. Again, availability of the reports might be one

explanation. Another reason, I speculate, can lie in the fact that different reports in the system are not quite well organized and there is no demonstration or function specification on the website to help people get started. Making the reports for teachers more user-friendly would also be of value. One teacher commented: “I got more familiar with the program as the year went on. It wasn't until the end of the year that I was able to start using the data. And that was only after I was shown step by step on how to retrieve the information and then use it.” Overall, the teachers seemed to value the ASSISTments system mostly for its tutoring capabilities rather than for its assessment capabilities.

I speculate that teachers that use the ASSISTments reports will learn more about their students, and thereby make their classrooms more effective, and thus produce better learning. However, we will not know if this is true until we have run a randomized controlled study with 20 teachers in the control group and 20 teachers in the experimental group.

3.4 Related Work and Contribution

Many researchers have been interested in constructing assessment/tutoring systems on different subjects, many of which provide the similar tutoring functionality as the ASSISTment system and various reports to teachers to help instructing student learning. Measures of Academic Progress (MAP -

http://www.nwea.org) are state-aligned computerized adaptive tests provided by the Northwest Evaluation Association (NWEA) and it is also the most commonly used assessment system by Worcester Public Schools. MAP covers subjects other than math and gives similar online reports such as class rosters, student progress report, class by subject report to educators to guide their instructions. Unlike the ASSISTments system, as an assessment system, MAP provides no tutoring to help student learning and it sticks to the strands and categorization given by the state. Hence, it lacks the ability to analyze a problem in further detail, and performance is reported in only 5 categories.

The Online Learning Initiative (OLI - http://www.cmu.edu/oli/) from Carnegie Mellon University provides a collection of online tutors directed at many subject areas. While the OLI provides a wide range of online tutors, the tutors lack extensibility to other tutor types and domains, resulting in a high cost for creating content.

Cognitive Tutors, created by LearnLab (http://www.learnlab.org/), also provide tutoring in addition to being extendable to other domain or content. They have been successful in raising students' math test scores in high school and middle-school classrooms. The Authoring tools, named, CTAT, are provided to make content creation easier for experts and possible for novices in cognitive science. However, although Cognitive Tutors trace knowledge mastery at a very

detailed level, they do not provide comprehensive reports about students' progress to teachers on the amount of assistance they need to finish the problems, etc. They do not give item analysis reports or class summary reports, either.

The National Center for Research on Evaluation, Standards and Student Testing (CRESST) (Vendlinski et al., 2005) provides an online system (not limited to math) and has a collection of tools to support the creation and distribution of content. However, the CRESST system does not offer tutoring, instead it allows for open ended questions that are then evaluated by a human teacher, nor does the CRESST system provides reports for teachers.

Effective Educational Technologies (EET) developed a series of online assessment and tutoring programs (MasteringPhysics - http://www.masteringphysics.com/, MasteringGeneralChemistry, and MasteringAstronomy) together with the authoring tools for content creation. Most like in the ASSISTments system, with the mastering program, students receive feedback based on common wrong answers and misconceptions. By capturing the step-by-step difficulties of individual students, the Mastering platform responds to each student with individualized hints and instructions. The program provides tools to find problems of the wanted type, topic coverage, and level (functioned as problem difficulty report as described in Section 4) and to monitor class/student performance via a "grade book"; tracks students' work on the sub-problems

(similar to the scaffolding questions in ASSISTments) and awards partially credit when evaluating students' performance. MasteringPhysics has been widely used as homework system while the ASSISTment project just started its first step into the picture.

LON-CAPA (http://www.lon-capa.org/) is a special assessment system because of its distributed learning content management that allows the sharing of assessment materials across institutions and disciplines. It provides assessment analysis gives an overview of how students are performing in the courses. The report shows all the attempts made by a student on each problem and it can also analyze one problem across all students, which is rather simple, comparing reports in the MAP, MasteringPhysics, or the ASSISTments system.

3.5 Conclusion

To summarize, I feel that I have developed some state-of-the-art online reporting tools that will help teachers and researchers be better informed about what their students know. The primitive evaluation is that we have made it possible for all these reports to work live in the classroom. Since doing learning analysis by hand is both time consuming and fallible, one aim of our reporting system is to automate statistical analysis of learning experiments. For instance, after researchers set up an

experiment, the system will send out emails automatically after the study is done and inform which condition is superior.

Chapter 4

4 Predicting Student Performance Better in a Tutoring System

Chapter organization: The first section of the chapter gives an introduction of the problem and describes how the later work is motivated. Section 2 focuses on developing dynamic assessment metrics to improve predictive power of assessment models. Section 3 shows how I have emphasized the problem of tracking student performance longitudinally. In Section 4, the dynamic assessment and longitudinal methods are combined to create a better prediction. Section 5 focuses on evaluating the estimates made from my models against those from standardized tests. Section 6 does a literature review of related works that are compared to my work to emphasize the contribution of this chapter. All studies from this chapter have been published in the following conference and journal papers: Feng, Heffernan, & Koedinger, 2006a, 2009 (Section 2, Section 3);

Feng, Heffernan, & Koedinger, 2006b (Section 4); Feng, Beck, Heffernan & Koedinger, 2008(Section 5).

4.1 Introduction

As mentioned in Chapter 1, critics of US government's No Child Left Behind legislation are calling the bill "No Child Left Untested". Among other things, critics point out that every hour spent assessing students is an hour lost from instruction. But, does it have to be? What if we better integrated assessment into classroom instruction and allowed students to learn during the test? In the ASSISTments system, we implement an approach that provides immediate tutoring on practice assessment items that students cannot solve on their own. The system helps teachers make better use of their time by offering instruction to students while providing a more detailed evaluation of student abilities to the teachers, which is impossible under current approaches.

Recently, in an interview with U.S. News & World Report (Ramírez, & Clark, 2009), U.S. Secretary of Education Arne Duncan weighed in on the NCLB Act, and called for continuous assessment. He mentioned that he is concerned about over-testing, and feels that fewer, better tests would be more effective. He wants to develop better data management systems that will help teachers track individual student progress in real-time, so that teachers and parents can assess

and monitor student strengths and weaknesses. To reflect on the "continuous assessment" idea, I think, one way that ITS research distinguishes itself from other educational software development is that it is concerned with modeling the knowledge of the learner in some computationally useful and inspectable way (McCalla & Greer, 1994). The modeling phase should involve understanding learner behavior in the rich context of the environment in which learning occurs, and, thus, obtains a better understanding of student's pre-existing, or current knowledge status and how the knowledge is changing over time.

ASSISTments provides an ideal test bed to investigate continuous assessment. It automatically provides students with feedback, scaffolding questions and hints based upon the answers given by students and their help-seeking request; it transparently and automatically logs all student actions into a background database; it keeps records of all student activities in the tutoring system over a long period of time; and, moreover, it synthesizes all related profile of students including demographic information, the school they go to, their teachers, and their state test scores over years. There is rich data about the nature and amount of help that the student was given, which would be of great value in judging a student's mastery of knowledge. Recall that I have presented our prediction of students' "expected" MCAS test scores as a single column in one of our online teacher reports, the "Grade Book" report, described in chapter 3. The predicted score was created based only upon student's average percent correct on the original

questions. In Razzaq et al. (2005), we reported a correlation between our prediction for the 68 students who have used our system in May 2004 and their real MCAS raw score (r = .7), which is somewhat satisfying given such a primitive model. But, can we do better than that? Examining the prediction model, I notice three issues that I will refer to as the *static* issue, the *average* issue, and the *uninformative* issue respectively. First, because the model only uses original question response data, it is static in that does not distinguish between two students, "Tom" and "Jack", who both got 39% of the original questions correct, but then needed very different levels of tutoring to get the problem correct eventually. "Tom" requested for 353 hints while "Jack" only asked for about half of that; in less time, "Jack" finished more problems than "Tom" did. Intuitively, "Jack" was probably at a higher proficiency level than "Tom" and, thus, should receive a higher predicted score. Secondly, the prediction is an average. It totally ignores the change of student's proficiency level in the past several months (If we assume the students used the system for 30 minutes every other week, then the 4 hours of practice has been accumulated in the past 4 months). Presumably, even if two students have the same average score in the middle of a year, the one who starts with a lower incoming knowledge but learns quicker probably will end up with a higher MCAS test score. Thirdly, the prediction is not informative for teacher's classroom instruction in that it does not tell teachers which particular knowledge component to be focused on.

In the following two sections, I argue the two issues have been addressed by leveraging the ample data resources in the ASSISTments system. Section 4.2 addresses the static issue by introducing dynamic metrics, while section 4.3 focuses primarily on the second issue, the average issue, and applies longitudinal models. I will leave the uninformative issue for Chapter 5 which is all about skills.

4.2 Addressing the Assessment Challenge Using Dynamic Metrics

The goal of this section is to see if it is possible to do assessment better if we take advantage of the student-system interaction information in ASSISTments that is normally not available in traditional practice tests. I will focus on answering the research question: *Does the tutoring interaction provide valuable assessment information?* A positive answer to the research question would help us to build a better predictive model and also improve our online teacher reporting. My hypothesis is that we can achieve more accurate assessment by not only using data on whether students get test items right or wrong, but by also using data on the effort required for students to solve a test item with instructional assistance. These measures include response efficiency that accounts for the time it takes students to come up with an answer to a problem, the time they take to correct an

answer if it is wrong, help-seeking behavior (e.g. the number of hints they request), and their performance on the sub-steps (called scaffolding questions).

4.2.1 Developing dynamic metrics

In order to differentiate student mastery at a finer grained level, I develop the following group of metrics that measure students' accuracy, speed, attempts and their help-seeking behavior:

- Original_Percent_Correct – students' percent correct on original questions, which we often referred to as the "static metric".
- Original_Count - the number of original items students have done. This measures students' attendance and how on-task they are. This measure also reflects students' knowledge since better students have a higher potential to finish more items in the same period of time.
- Scaffold_Percent_Correct - students' percent correct on scaffolding questions. In addition to original items, students' performance on scaffolding questions is also a reasonable reflection of their knowledge. For instance, two students who get the same original item wrong may, in fact, have different knowledge levels and this may be reflected in the one may get more of the scaffolding questions right than the other.

- Question_Count - the number of questions (both original items and scaffolding questions) students have finished. Similar to Original_Count, this variable is also a measure of attendance and knowledge but given the fact that scaffolding questions show up only if students fail the original question, it is not obvious how this measure will correlate with students' MCAS scores.
- Hint_Request_Count - how many times students have asked for hints.
- Avg_Hint_Request - the average number of hint requests per question.
- Bottom-Out_Hint_Count - the total number of bottom-out[9] hint messages students got.
- Avg_Bottom_Hint - the average number of bottom-out hint messages students got per question.
- AvgFirstHintRequest_original – on average, how many times students requested hints before they made an attempt on an original question.
- AvgFirstHintRequest_scaffold – Similar to AvgFirstHintRequest_original, this variable measures on average, how many times students requested hints before they made an attempt on scaffolding questions.

9 Since the ASSISTment system does not allow students to skip problems, to prevent students from being stuck, most questions in the system were built such that the last hint message almost always reveals the correct answer. This message is referred to as "Bottom-out" hint.

AvgFirstHintRequest_original and AvgFirstHintRequest_scaffold are only available for 2005-06 data. I hypothesize the two metrics are both a reflection of student knowledge and their confidence. Students who either do not understand the question or are uncertain of their response may request hints before making an attempt. Accordingly, my hypothesis is that both metrics will be negatively correlated with MCAS scores without other variables in the model. If Original_Percent_Correct is already in the model, then AvgFirstHintRequest_original might be positively correlated because it may be better to know you don't know (and those ask for a hint) than not know and just guess.

- Attempt_Count - the total number of attempts students made across all original and scaffolding questions.
- Avg_Attempt - the average number of attempts students made for each question.
- Avg_Item_Time - on average, how long it takes for students to finish a problem (including all scaffolding questions if students answered the original questions incorrectly).
- Avg_Question_Time - on average, how long it takes for a student to answer a question, whether original or scaffolding, measured in seconds.

- Total_Minutes - how many total minutes students have been working on items in the ASSISTment system. Just like Original_Count, this metric is an indicator of the attendance.

The ten measures from Hint_Request_Count to Avg_Question_Time in the above list are generally all ASSISTment style metrics (or the assistance metrics), which indicate the amount of assistance students need to finish problems and the amount of time they spend to finish items. Therefore, the hypothesis is that these measures will generally be negatively correlated with MCAS scores (though, for instance, Attempt_Count or Total_Minutes may be positively correlated because they are partially determined by time on task).

Among these measures, "Original_Percent_Correct" is a static metric that mimics paper practice tests by scoring students as either correct or incorrect on each item, while the rest of the measures are dynamic testing metrics that measure the amount of assistance students need before they get an item correct, how fast they answered the questions, etc.

I have been presenting some of these online measures in our reports to teachers (See Figure 3.1). Particularly, student "Mary" used the system 4 hours and 17 minutes, finished 114 items with 20% correct. She went through 356 scaffolding questions with 20% correct and asked for 705 hints, which is excessive compared to her classmates.

4.2.2 Description of the data

The first data I consider comes from the 2004 – 2005 school year, the first full year in which the ASSISTment system was used in classes in two middle schools in Massachusetts. At that time, the ASSISTment system contained a total of 493 main questions and 1216 scaffolds; 912 8th grade students' logs were maintained in the system over the time period from September to May. Of these, approximately 400 original questions and their corresponding scaffolding questions were in regular use by students. Although the system was web-based and hence accessible in principle anywhere/anytime, students typically interacted with the system during one class period in the schools' computer labs every few weeks. Among these 912 students, I am able to obtain complete data for 417 of them. The data set contains online interaction data from the ASSISTment system, the results of 8th grade MCAS tests taken in May, 2005 and the results of 10th grade MCAS tests taken by the same group of students two years later, in May, 2007. I exclude the data of 25 students who did less than 39[10] questions in ASSISTments. The 391 students in the final 04-05 data set have practiced mathematics problems in the ASSISTment system for a mean length of 267

[10] The number 39 was picked because there are 39 questions in the real 8th grade MCAS test each year.

minutes (standard deviation = 79) across about 9 sessions, finishing on average 147 items (standard deviation = 60).

The second data set I will use is from the 2005-2006 school year. About 3,000 students used the system during the year, among which I collected a full data set for 616 students from Worcester Public Schools, including the online data from ASSISTments and their 8th grade MCAS test raw scores (MCAS test was taken in May, 2006). The students on average worked in the ASSISTment system for 196 minutes (standard deviation = 76), and finished an average of 88 items (at least 39 items, standard deviation = 42).

In the reporting system, student performance is represented in terms of scaled score and performance level (as seen in Figure 3.1) because teachers and students are more used to interpreting these metrics than understanding the raw score. But since the raw score provides a finer grain differentiation between students, when I analyze the data, all of the procedures are judged on their ability to predict the MCAS raw scores.

Given the data set, my goal is to see if we can reliably predict students' MCAS raw scores and to evaluate how well on-line use of the ASSISTment System, can help in the prediction. To achieve the goal, I perform a stepwise linear regression using the online measures as independent variables to predict students' MCAS scores. I first apply the approach on the 2004-05 data and then on

the 2005-06 data. Models constructed based on data from different years are also cross-validated too investigate whether the approach generalizes across years.

4.2.3 Modeling

Now that the dynamic metrics are developed, I first present the Pearson correlations between actual MCAS raw scores and all the independent variables in Table 4.1 to give some idea of how these factors are directly related to MCAS score. The first column shows the correlation between the factors and the 2005 MCAS test scores in the 2004-05 data set, and the second column shows the correlation between the factors and the 2006 MCAS test scores. All these factors except Attempt_Count and Avg_Question_Time turn out to be significantly correlated with MCAS score at the 0.01 level (2-tailed). In general, students who maintained higher percent correct, finished more items, requested for less help and solved problems faster in ASSISTment, tended to have a high MCAS score at the end of the year, which is consistent with my hypothesis.

Table 4.1. Correlations

Correlations		Factors	04-05 data/ 2005 MCAS (391 students)	05-06 data/ 2006 MCAS (616 students)
Online Computer Metrics	Static metric	Original_Percent_Correct	0.786	0.689
	Attendance	Original_Count	0.477	0.394
		Total_Minutes	0.241	0.202
	Assistance Style Metrics (dynamic metrics)	Scaffold_Percent_Correct	0.721	0.647
		Question_Count	0.163	0.132
		Hint_Request_Count	-0.465	-0.265
		Avg_Hint_Request	-0.689	-0.536

Correlations		Factors	04-05 data/ 2005 MCAS (391 students)	05-06 data/ 2006 MCAS (616 students)
		Bottom_Out_Hint_Count	-0.423	-0.258
		Avg_Bottom_Hint	-0.584	-0.477
		Attempt_Count	**0.036**	**0.026**
		Avg_Attempt	-0.398	-0.593
		Avg_Question_Time	**-0.062**	**-0.026**
		Avg_Item_Time	-0.391	-0.299
		AvgFirstHintRequest_original	NA	-0.347
		AvgFirstHintRequest_scaffold	NA	-0.351

My effort to predict student MCAS scores starts with **a "lean" model**, an Item Response Theory (IRT)-style model using ASSISTments data only. As a starting point, I use a one-parameter IRT model (the Rasch model, also called 1-PL IRT model) (van der Linden & Hamilton, 1997), the straightforward model with just student proficiency and item difficulty parameters (and only on original question data). In the simple Rasch model, the probability of a correct response is modeled as a logistic function of the difference between the student and item parameter. To train the lean model, I use a data set of all data collected in the ASSISTment system from Sept., 2004 to Jan., 2008, including responses to 2,797 items (only on original questions) from 14,274 students. By including more student response data from the four years, I hope to acquire a more reliable estimate of the item parameters, and thus a more stable estimate of the student proficiency.

Some readers may argue for multi-parameter models, such as 2-PL or 3-PL IRT models, that are expected to perform better on parameter estimation given sufficient amount of data. There were extensive research with regard to the test length and sample size on parameter estimates for dichotomous items in the 1980s. For instance, Hulin, Lissak, and Drasgrow (1982) concluded that 30 items and 500 examinees were required to produce acceptable results when the 2-PL model was used. For the 3-PL model, they suggested 60 items and 1,000 examinees. We have thought of using more general models. However, the number of data points per item in our data set varies tremendously (mean = 276, and standard deviation = 463). There are 424 items for which we have less than 10 data points. So, we stick with the simplest 1-PL Rasch model.

I fit the Rasch model in BILOG-MG 3.0 (Zimowski, Muraki, Mislevy & Bock, 2005) and obtain an estimate of every student's proficiency trait score (usually called the student parameter), ranging from -4 to 4. The trait score is then transformed to the probability of a correct response on an item of average difficulty (i.e. item difficulty parameter equal to 0.5). I name this probability "IRT_Proficiency_Estimate" and add an extra column in the data set that is described in section 4.2.1 to record the student proficiency estimates. Finally, I conduct a linear regression to predict MCAS scores using student proficiency estimate to put them on the same scale.

The lean model is based on how students performed on the original questions but ignores how students interacted with the system. Using the online testing metrics, I run a stepwise regression[11] analysis to predict MCAS test scores. I will refer to this model as **the assistance model.** Original_Percent_Correct is not included in the assistance model in order to distinguish it from the lean model and to show how the interaction information alone can do in estimating student performance. It is worth highlighting that the assistance model, which does not use the assessment information from the original questions is fundamentally different from traditional assessment model, represented by the lean model which only uses the original question information. By contrasting the assistance model against the lean model, we can isolate the assessment value of the tutor-student interaction session.

A third model that combines the student proficiency parameters, estimated by the lean model, with the online testing metrics is explored as well. New parameter values for the assistance metrics are again fit in a stepwise regression. I name this model **the mixed model**. Student percent correct on original questions is not included here because these values have been used to find the best fitting student

[11] We set probability of F to enter <= .05; probability of F to remove >= .10 for all stepwise regression analysis we did.

proficiency parameters. Using student proficiency has the advantage that it takes into consideration the difficulty of each item a student attempts.

As mentioned before, Beck, Jia & Mostow (2004) were able to increase the within-grade correlation between their model and student performance on a fluency test significantly by adding help-seeking behavior in the computer Reading Tutor. In order to see how much the information on the amount of assistance that students need to solve a problem can help in the ASSISTment system, I construct another model, which I will refer to as **the help model**. In Beck et al. (2004), two features were constructed to describe student help requests: the percentage of words on which the student clicked for help and the percentage of sentences on which the student requested help. While in this dissertation, I have developed more features, including Hint_Request_Count, Avg_Hint_Request, Bottom_Out_Hint_Count, Avg_Bottom_Hint. Recall that all these hint related metrics are significantly correlated with MCAS test scores. Thus, to build the help model, I run a stepwise linear regression to predict student MCAS scores using the student proficiency parameter estimated by the lean model plus all of the hint related metrics.

So far I have constructed four models: the lean model, the assistance model, the mixed model and the help model. Each model addresses different aspects of student performance. The modeling process has been replicated using data from

both 2004-2005 and 2005-2006 school years. In the next section, I will evaluate the models.

4.2.4 Results and model evaluation

4.2.4.1 Do the dynamic metrics help build a more predictive model?

In Table 4.2, I summarize the four models that have been built using 2004-2005 data, for which I have selected different groups of independent variables (IV) for regression. For each model, the table report how many variables entered into the model, how well the model fit data and how the predicted scores correlated with the real MCAS scores. SPSS automatically displays R Square and Adjusted R Square in the output. Because the four models have different numbers of parameters and are not always nested[12], we need a method to compare the generalization quality of the models. I choose to use the Bayesian Information Criterion (BIC) for this purpose and applied the formula for linear regression models introduced by Raftery (1995, p135), which is different from what is typical used for calculating BIC but most convenient for linear regression models:

$$BIC = n\log(1-R^2) + p\log(n)$$

where

[12] Two models are nested if they both contain the same terms and one has at least one additional term.

n: the sample size (for the 2004-2005 data case, *n* = 391; for the 2005-2006 data, n=616)

log: natural logarithm

p: the number of independent variables included in each model (not including intercept)

The MCAS test score predictions of the best fitting models are compared with actual 2005 MCAS test scores to calculate mean absolute deviation (MAD), which has been used as the measure to evaluate the student models in prior works (e.g. Anozie & Junker, 2006; Ayers & Junker, 2006; Feng, Beck, Heffernan & Koedinger, 2008). Here I follow those works and compute MAD for the models. The measure of MAD is suggested by Brian Junker, a statistician from Carnegie Mellon University (Junker, 2007).

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

where $MCAS_i$ is the actual MCAS score of the i_{th} student, and $prediction_i$ is the predicted score from the prediction function being evaluated.

The correlation between the adjusted predicted scores[13] by each model and student real MCAS test scores is also presented. The absolute deviations (AD) of each model for each of the 391 students are compared with each other, two models at a time, using paired t-tests (the *p*-values for comparisons of successively better models are shown in the first column of Table 4.2).

Table 4.2. Model Summary (based on 2004-2005 data)

Model	MAD	R Square	BIC	#variables	Correlation with 2005 8th grade MCAS
The lean model	6.40	0.537	-295	1	0.733
p=.031					
The help model	6.13	0.585	-326	3	0.765
p=.004					
The assistance model	5.46	0.674	-402	6	0.821
p=.001					
The mixed model	5.04	0.707	-450	5	0.841

As we can see, the help model correlates reliably better with MCAS score than the lean model, suggesting that in ASSISTments, student help-seeking behavior is helpful in improving MCAS score prediction. The lean model does not correlate with MCAS score as well as the assistance model, which indicates that MCAS score can be better predicted by using features reflecting student assistance requirement, effort, attendance, etc, even if we ignore student responses to original questions. Additionally, we can improve our prediction of MCAS score further on top of the assistance model by combining the student proficiency

[13] The adjusted predicted score is calculated by doing "leave-one-out" cross validation in SPSS 14.0.

parameter together with the online metrics that describe how students interacted with the system, such as the number of attempts students need, and how long a student need to answer a question. The improvement is statistically reliable in terms of MAD[14], BIC[15] and the correlations[16].

As the best fitted model, the mixed model takes into account students' performance on both original items (indicated by the IRT_Proficiency_Estimate) and scaffolding questions (indicated by the feature Scaffold_Percent_Correct), together with the other online metrics, and the model correlates with MCAS scores fairly well (r = 0.841). With five variables (one less than in the assistance model), it gains a 0.033 increment in the R^2 value, 48 points lower on BIC value, and a significant better correlation with MCAS scores, which means the mixed model is significantly better than other models. Variables enters the mixed model in the following order: IRT_Proficiency_Estimate, Scaffold_Percent_Correct, Avg_Question_Time, Avg_Attempt and Avg_Hint_Request. Among these variables, IRT_Proficiency_Estimate and Scaffold_Percent_Correct have positive coefficients, and Avg_Question_Time, Avg_Attempt and Avg_Hint_Request have negative correlation coefficients. Notice that student percent correct on

[14] See the chapter of "Evaluating hypotheses" in Mitchell (1997).

[15] Raftery (1995) discussed a Bayesian model selection procedure, in which the author proposed the heuristic of a BIC difference of 10 is about the same as getting a p-value of p = 0.05.

[16] We did significance testing of the difference between two dependent correlations from a single sample using the applet online at http://www.quantitativeskills.com/sisa/statistics/correl.htm.

original questions (IRT_Proficiency_Estimate) is considered the most significant predictor since it entered the final model earlier than other factors. However, the information from the tutoring session is helpful too, which indicates that performance on sub-steps, the speed, and the help-seeking behavior are additional, reliable indicators of students' level of knowledge acquisition. The mixed model is presented in Table 4.3 and the interpretation of the model is straightforward:

- Every point increase in IRT_Proficiency_Estimate (ranging from -4 to 4) adds 26.800 points to the prediction of MCAS score.
- Every one percent increase in Scaffold_Percent_Correct (ranging from 0 to 1) adds 20.427 points to the prediction of MCAS score.
- Average students' predicted score will reduce 0.170 points for every extra second students spent to finish a question.
- On average, if a student needs one more attempt to reach a correct answer for an item, his/her predicted MCAS score will be 10.5 points lower.
- On average, if a student requests one more hint, the predicted score will lower 3.217 points.

Table 4.3. Variables and coefficients of the mixed model (based on 2004-2005 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	32.414		6.136	0.000
1	IRT_Proficiency_Estimate	26.800	0.443	10.364	0.000
2	Scaffold_Percent_Correct	20.427	0.283	4.436	0.000

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
3	Avg_Question_Time	-0.170	-0.212	-6.941	0.000
4	Avg_Attempt	-10.500	-0.178	-5.485	0.000
5	Avg_Hint_Request	-3.217	-0.149	-2.175	0.030

The variables and corresponding coefficients of the final help model and the final assistance model are shown in Table 4.4 and Table 4.5. As we can see in Table 4.4, in addition to the IRT_Proficiency_Estimate, the average number of hint requests per question (Avg_Hint_Request) and the total number of hint requests (Hint_Request_Count) are also significant predictors of student performance on the MCAS test. Comparing Table 4.5 with Table 4.3, we notice that the assistance model is not nested within the mixed model. Without the IRT_proficiency_estimate, the step-wise regression added two other variables to the assistance model: Original_Count (total number of problems finished) and Attempt_Count (total number of attempts).

Table 4.4. Variables and coefficients of the help model (based on 2004-2005 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	17.021		6.816	0.000
1	IRT_Proficiency_Estimate	31.389	0.519	10.430	0.000
2	Avg_Hint_Request	-8.967	-0.416	-7.019	0.000
3	Hint_Request_Count	0.008	0.160	3.284	0.001

Table 4.5. Variables and coefficients of the assistance model (based on 2004-2005 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	39.698		7.124	0.000
1	Scaffold_Percent_Correct	16.383	0.227	3.348	0.001
2	Original_Count	0.114	0.572	7.775	0.000
3	Attempt_Count	-0.029	-0.473	-6.712	0.000

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
4	Avg_Question_Time	-0.131	-0.162	-3.941	0.000
5	Avg_Hint_Request	-6.933	-0.322	-4.774	0.000
6	Avg_Attempt	-5.349	-0.091	-2.236	0.026

I replicate the same modeling process on the data from the school year of 2005-2006. As mentioned before, the data is for 616 8th grade students from Worcester Public Schools. All of the four models are fitted to this data set and the results are summarized in Table 4.6.

Table 4.6. Model Summary (based on 2005-2006 data)

Model	MAD	R Square	BIC	#variables	Correlation with 2006 8th grade MCAS
The lean model	5.77	0.615	-581	1	0.784
The help model ($p<.001$)	5.46	0.669	-656	4	0.818
The assistance model ($p=.693$)	5.39	0.666	-630	7	0.816
The mixed model ($p<.001$)	4.89	0.728	-763	6	0.853

The results are similar to those obtained using the 2004-2005 data set where the mixed model is still the best fitting and most generalizable model. The adjusted predicted score correlated with the 2006 MCAS test scores at 0.853, which is reliably higher than the correlation between the other models and the MCAS scores. The ranking of the other three models is the same as the ranking we got from using the 2004-2005 data except that the difference between the help model and the assistance model is no longer reliable (p = 0.693). The assistance

model and the mixed model are still reliably better than the lean model ($p < 0.001$).

A critic may argue that it is not fair to have the lean model as a contrast case as students were not spending all their time on assessment. Whether or not the ASSISTment system would yield better predictions than such a tougher contrast case, where students only spend on-line time on assessment and not on instruction, is an open question worthy of further research. However, we would remind that critic that such a contrast would leave out the instructional benefit of the ASSISTment system and, moreover, might not be as well received by teachers and students. Schools are willing to use the ASSISTments often because they believe (and Razzaq et al., 2005 have shown) students learn during the ASSISTments. However, the schools are not willing to use more valuable instruction time to test more often.

Some readers may have noticed that no quadratic terms or interactions between factors are included in our models when building regression models. As a matter of fact, I have suspected that there might be a non-linear relationship between the online measures and MCAS scores and such a regression model is also trained. I got a much more complicated model; the models with quadratic and interaction terms fit (by BIC values) the best on training data of both years, yet they are not statistically reliably better than the mixed models where no quadratic

or interaction terms are included. With 25 variables entered (mostly interaction terms), the model becomes hard to interpret and potentially overfits[17] our data. Both for clarity and because our goal is generalization (e.g., better cross validation and BIC values) not just fit, the mixed model is our preferred model for MCAS score prediction.

4.2.4.2 Model validation

Now that I have trained models on data from two different years, I want to compare the models and see how different they are from each other. The coefficients of the mixed model trained based on 2005-2006 data is shown in Table 4.7. Comparing the model here with the model trained using 2004-2005 data as shown in Table 4.3, we can notice that the model is entered with 6 variables, yet the mixed model trained using 2004-2005 data is more parsimonious (5 variables). Further, it is worth pointing out that the assistance metrics that measure student help-seeking behaviors does not enter the model as they do for the 04-05 model (shown in Table 4.3). In both models, the student proficiency parameter estimated by the IRT model and student percent correct on scaffolding questions are the top two predictors.

[17] We applied the model trained based on 2004-2005 data on the data from 2005-2006 and the predicted value does not correlate with 2006 MCAS score very well. The correlation is at the same level as the lean model.

Table 4.7. Variables and coefficients of the mixed model (based on 2005-2006 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	3.284		2.224	0.027
1	IRT_Proficiency_Estimate	32.944	0.530	13.657	0.000
2	Scaffold_Percent_Correct	21.327	0.309	8.544	0.000
3	Question_Count	0.072	0.652	8.891	0.000
4	Avg_Question_Time	-0.102	-0.173	-5.498	0.000
5	Avg_Item_Time	0.045	0.154	5.432	0.000
6	Total_Attempt	-0.044	-0.550	-7.570	0.000

Since the models are not the same across years, the models might be overfitting the training data. Presenting BIC that penalizes a model by the number of variables entering the model and using adjusted predicted score that is calculated using "leave-one-out" cross validation have given us some protection against overfitting. On top of that, I investigate the issue further and explore how well the models generalize by validating the model using data in the same year and across years, and then evaluate the model fitting using the correlation and the MAD on the testing data. The first thing is cross validating the model (the mixed model) using the same year's data. For this purpose, I allow SPSS to randomly select 50% of the 2004-2005 data as the training cases and trained the mixed model on the 203[18] selected training cases. It is good to see that the mixed model correlated with the 2005 MCAS score with r equal to .842 ($p < .01$). The resulting model is also very similar to the one shown in Table 4.3, where the same

[18] Sharp readers may notice 203 is not exactly 50% of 391, the total number of students in the data set. The fact is, SPSS finished the 50% sampling and returned 203 rows, and we decided not to change it.

variables enter the model and the sign of all the coefficients are in the same direction as before, except that the values of the estimated coefficients changed slightly. I then fit the model on the 188 testing cases of the same year, and find out that the model fit well on the testing data (r = .837, MAD = 5.25). In the next step, I train the model using the second half of the data and test it using the first half, and get similar result. Thus, it is verified that the mixed model worked very well on the data of the same school year (i.e. the same group of students, and the same MCAS tests, etc.)

Now that I have validated our model inside a year, I further test its validity using data across years. I use the 2005-2006 data as the testing set for the model trained over the 2004-2005 data and vice versa[19]. Again, I do the validation on the mixed model as it is the preferred model. It turns out that both models work fairly well on the testing set. As presented in Table 4.8, the predicted scores tested on the 2005-2006 data correlate with the 2006 MCAS with the correlation coefficient equal to 0.827 and the correlation is 0.824 for the model tested on 2004-2005 data. And both correlations are significant at the 0.01 level. Therefore, even though the models constructed based on data from different years are not quite the same, they

[19] Since the range of MCAS raw score is 0 to 54, if the model predicted a score larger than 54 on the testing data, we only assigned a score of 54; and if the model predicted a negative score, we assigned a score of zero.

are both very predictive of student end-of-year exam scores. The strong correlations do a good job of protecting us from overfitting the data.

Table 4.8. Results of testing the mixed models on a different year's data

Year of training data	Year of Testing data	MAD on Testing data	Correlation with MCAS scores on testing data	MAD after correction
2004-2005	2005-2006	5.82	0.827	5.16
2005-2006	2004-2005	6.55	0.824	5.80

However, I also notice that the MADs on testing data sets are relatively high. Given that the correlations are solid, I suspect that the models are overall under-predicting or over-predicting on the testing data sets. For this reason, I generate the scatter plots for both testing sets (shown in Figure 4.1). As we can see from the scatter plots (the fit line is shown), the model trained over 2005-2006 data under-predicts the 2005 MCAS scores for the students who used the ASSISTment system during the school year 2004-2005; while the model trained over the 2004-2005 data over-predicts the 2006 MCAS score for our users during 2005-2006. Why? Presumably, if our models are really overfitting, we probably would not get a solid correlation; and since the IRT_Proficiency_Estimate in the mixed model can account for student difference, we speculate that if the difficulty of the MCAS tests shifted in the two years that would directly have an impact on student raw

scores[20] though their performance level wouldn't be affected because of the equalization procedure as described later.

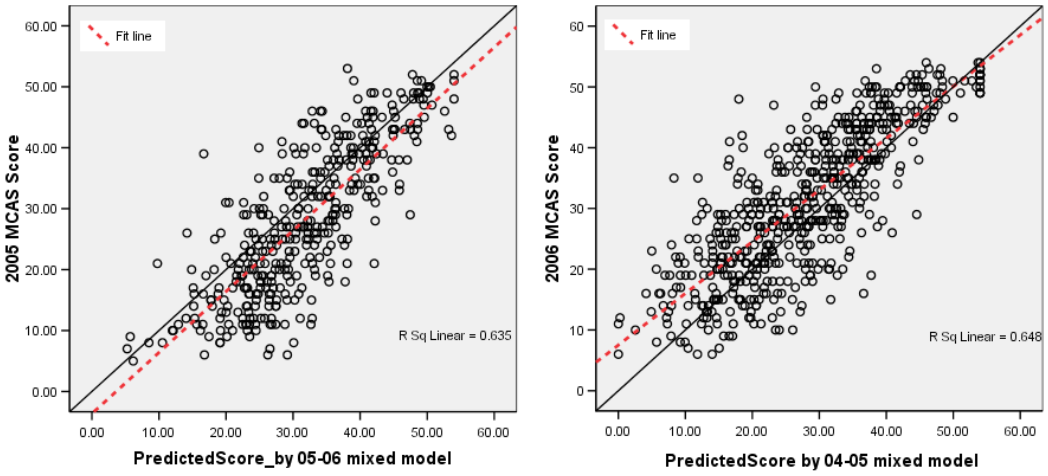


Figure 4.1. Scatter plot of predicted scores on testing data vs. MCAS raw scores

To investigate whether the difficulty of the MCAS tests changed from the year 2005 to 2006, I compute the average percent correct of all the students from Worcester for both tests. As I have suspected, overall, the students from Worcester got 42% correct on the 2005 MCAS test and 46% correct on the 2006 MCAS test which is statistically reliably higher than that of the 2005 test suggested by the two sample t-test of the two groups of students ($p < .05$). This indicates that the 2006 MCAS test was reliably easier than the 2005 MCAS test

[20] Recall that we have been using MCAS raw scores as the dependent variables in our models which are affected by the variant in the difficulty of the tests.

for students from Worcester. The result was confirmed by the fact that the threshold raw scores shifted 3 points higher in the year 2006 – students were required to score higher in 2006 to achieve the same performance level that their score would have received in 2005. This finding can be considered as an explanation of why the 2006 MCAS scores are under-predicted by the 2004-2005 mixed model while the 2005-2006 model over-predicts the 2005 MCAS score. Given this, I apply a very simple correction on the predicted score by subtracting 3 points from the scores predicted by the 2005-2006 model and adding the same to the scores predicted by the 2004-2005 model. This simple correction lowered the MADs by 0.7 points in both cases. Namely, the MAD drops from 5.89 to 5.16 when I test the on 2005-2006 data, and MAD drops from 6.49 to 5.80 when I use 2004-2005 data as the testing set.

To summarize, the models work fairly well on both the same year's data and data from different years. There are a number of possible reasons why the regression modeling process picked up different groups of predictors for data logged in different years including random variation and differences in the logging system, the user interface and the problem sets from one year to the next. Nevertheless, both models generalized and fit well on the testing set with some offsets caused by the changing of the MCAS tests themselves. With regard to the validation result, I claim that the dynamic testing approach can reliably improve the prediction of student end-of-year test scores by taking into consideration the

help-seeking behavior, the speed and the number of attempts needed to answer questions.

4.3 Tracking Learning Longitudinally

In Razzaq et al. (2005) and Feng, Heffernan, Beck & Koedinger (2008), we reported results that suggested students were learning directly during the assisting in ASSISTments. We did this by looking at groups of items that had the same skills and looked to see if performance later in the class period was associated with high performance. The gain score over all of the learning opportunity pairs suggested that students were learning in the system. In this section, instead of discussing within-system learning, I will be focusing on tracking student progress that results from both classroom instruction and ASSISTments tutoring over a long period of time (Feng, Heffernan, & Koedinger, 2006a). I am interested in characterize each student's pattern of change over time as well as if there is inter-individual differences in the change; and if the answer is yes, what the factors that are associated with the differences are. To investigate these questions, I conduct longitudinal analysis (Singer & Willett, 2003; Fitzmaurice, Laird & Ware, 2004) by fitting mixed-effect models on the ASSISTment data to investigate if learning happens over time. Specifically, the research questions I want to answer in this section include: *(a) Can we track student learning over the course of the year? (b)*

Can we see what factors affect student learning? (c) Can we track the learning of individual skills? I build a series of mixed-effects models and gradually introduce factors such as what school they are in, who their teacher is, or which class they are from into the models. I will be focusing on the first two questions in this section while leave the discussion on the third question to Chapter 4 that all has to do with skills: the construction of skill models, using skill models in tracking student performance, and evaluation of the skill models.

4.3.1 Longitudinal data analysis and mixed-effects models

"Singer and Willett" (Singer & Willett, 2003) style longitudinal data analysis is an approach for investigating change over time; in our case, the change of students' performance over the course year. Since students' typically use ASSISTments regularly over a long period of time, our data naturally fit into the area of longitudinal analysis. With that said, there are three methodological features that we need to check to make sure the data is suited to the analysis of change. First, each subject should have *three or more waves of data* (i.e. three or more observations at different measurement occasions). Two waves of data cannot describe the process of change since it cannot tell the shape of individual growth trajectory, nor can it distinguish true change from measurement error. Secondly, since we are talking about change, *time* is the fundamental predictor, and thus should be measured in a sensible metric. Theoretically, it could be measured in a variety of

units - years, semesters, months, days, and so on. Yet, Singer & Willett (2003) also warned that the metric should also be carefully selected so that it is useful for the outcome, and the spacing of the waves of data is appropriate for a particular study. Thirdly, we need to choose a *continuous outcome* that changes systematically over time. Statistical models care little about the meaning of the individual outcomes. Yet, in order to represent change trajectories using a meaningful parametric form, the outcome should be chose in a way such that it makes conceptual and theoretical sense for the outcome to follow such a trajectory.

Longitudinal data analysis allows us to learn a slope that represents a student's learning rate and intercept that represents the estimate of incoming knowledge for each individual student. This is achieved by fitting a multilevel statistical model (also referred to as mixed-effects model) as illustrated in more detail below. A two level mixed-effect model embodies two types of research questions. Level-1 questions about within-person change – How does each subject change over time?, and it's the responsibility of level-1 submodel to describe how individuals change over time; while level-2 questions about between-person differences – What predicts differences among people in their changes? And the level-2 submodel deals with systematic interindividual differences in change. Precisely, a multilevel (2 levels here) statistical model for longitudinal analysis can be represented as:

Level-1 submodel (individual growth model)

$$Y_{ij} = \pi_{0i} + \pi_{1i} * TIME_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0 \quad , \sigma_{\varepsilon}^{2}\right)$$

Level-2 submodel

$$\pi_{0i} = \gamma_{00} + \gamma_{01} * LEVEL_i + \varsigma_{0i}$$
$$\pi_{1i} = \gamma_{10} + \gamma_{11} * LEVEL_i + \varsigma_{1i} \qquad \begin{bmatrix} \varsigma_{0i} \\ \varsigma_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

Composite model

$$Y_{ij} = \gamma_{00} + \gamma_{01} LEVEL_i + \gamma_{10} * TIME_{ij} + \gamma_{11} * LEVEL_i * TIME_{ij} + (\varepsilon_{ij} + \varsigma_{0i} + \varsigma_{1i} * TIME_{ij})$$

where Y_{ij} is the score for person *i* at time *j*, a linear function of *TIMEij (CenteredMonth here);*

π_{0i}, π_{1i} are intercept and slope of change trajectory of subject *i* (i.e. π_{0i} represents student *i*'s initial knowledge, and π_{1i} indicates how much his/her learning rate)

γ_{01}, γ_{10} represents population average of level-1 intercept/slope

γ_{01}, γ_{11} are population average difference in π_{0i}, π_{1i} for a unit difference in level-2 predictor ($LEVEL_i$)

ε_{ij} is the random measurement error for person *i* at occasion *j*

ς_{0i}, ς_{1i} are parameter residuals that permit the level-1 parameters of one person to differ stochastically from those of others.

4.3.2 Description of the data

I use ASSISTments log data from 2004-2005 school year for the longitudinal analysis. Instead of formatting the data as one row for each student (often referred to as "person-level" data set), I choose to build a data set that had one record per student for each time they used the ASSISTments system (usually referred to as "person-period" data set, Singer & Willett, 2003). Each row represents a measurement occasion, i.e. a day when they came to the lab. Rather than treating time as days from the beginning of the year, I collapse All data in a given month is collapsed and month is chose as the level of granularity to measure time in. The time metric is called "CenteredMonth" since the months are centered around September 2004. Data from September, 2004 is excluded to account for the fact that students were learning how to use the system when they first came to the labs in September, 2004. This makes CenteredMonth run from 1 for Oct., 2004 to 9 for June, 2005. The final "person-period" structured data set contains 841 students and on average 5.7 data waves for each of them. This data set has more students than the one from above because we do not need to have student's MCAS test scores, nor is there a limitation on the total number of problems they need to finish. The data set contains data from students of the same 8 teachers from 2 schools, teaching a combined total of 31 mathematics classes. Students' percent correct on the original items is treated as the outcome whose values change systematically over months. To mimic the real MCAS score, we multiply percent

correct by 54 (the full MCAS score). This makes the range of the outcome change to 0~54. The outcome will be referred to as *PredictedScore* from now on.

4.3.3 Modeling

Most people would think the answer to research question (a) is obviously, yes, which is just what I thought before peeking into our data. In Figure 4.2, average %correct of students is plotted versus time. The y-axis is the average percent correct on the original questions (student performance on the scaffolding questions is ignored in this analysis). The x-axis represents the CenteredMonth (i.e. the number of months elapsed since the beginning of the year, Sept., 2004), where data is bunched together into months, so some students who came to the lab twice in a month will have their numbers averaged. Does it look like there is learning happening over time? Well, maybe, but not so “obvious”.

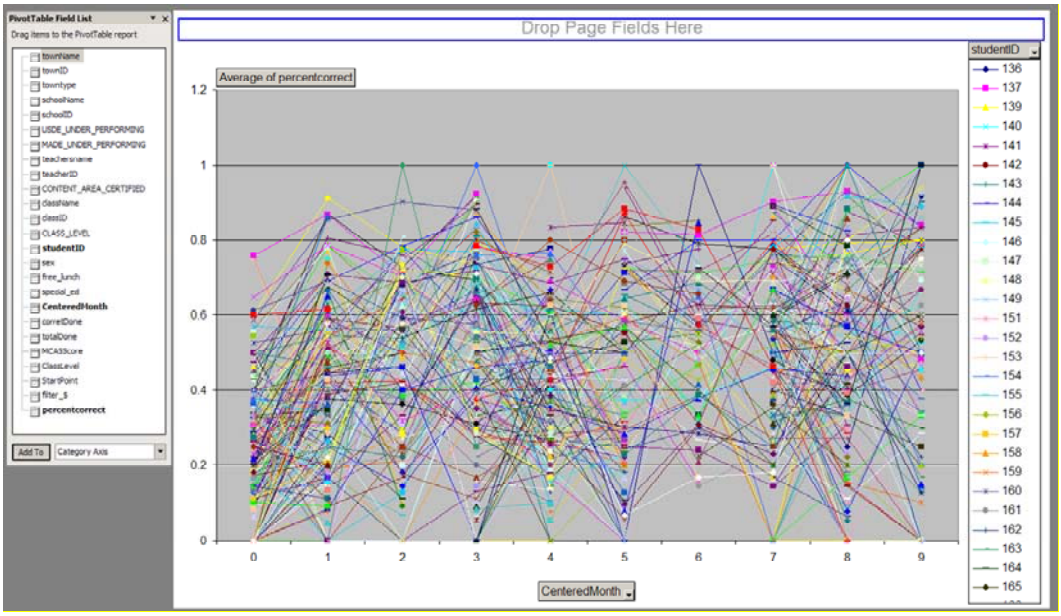


Figure 4.2. Average student performance is plotted over time

I plot data and fit a regression line for each individual student (Figure 4.3) before applying a mixed effect model and ended with a mean slope of 1.46 points per month. This positive trajectory slope is considered as the initial evidence that students learned across time. I then plan to build a series of mixed effect models of increasing complexity that would progressively add components for tracking learning over time, and then add parameters for schools, teacher and classes as shown in Figure X. I follow the standard method of performing longitudinal data analysis (Singer and Willett, 2003) fitting *mixed* effects (fixed effect plus random effect) models that simultaneously builds two sub-models, in which level-1 sub-model fits *within-person* change and describes how individuals change over time

and level-2 sub-model tracks *between-person* change and describes how these changes vary across individuals. The models are fitted in SPSS 13.0.

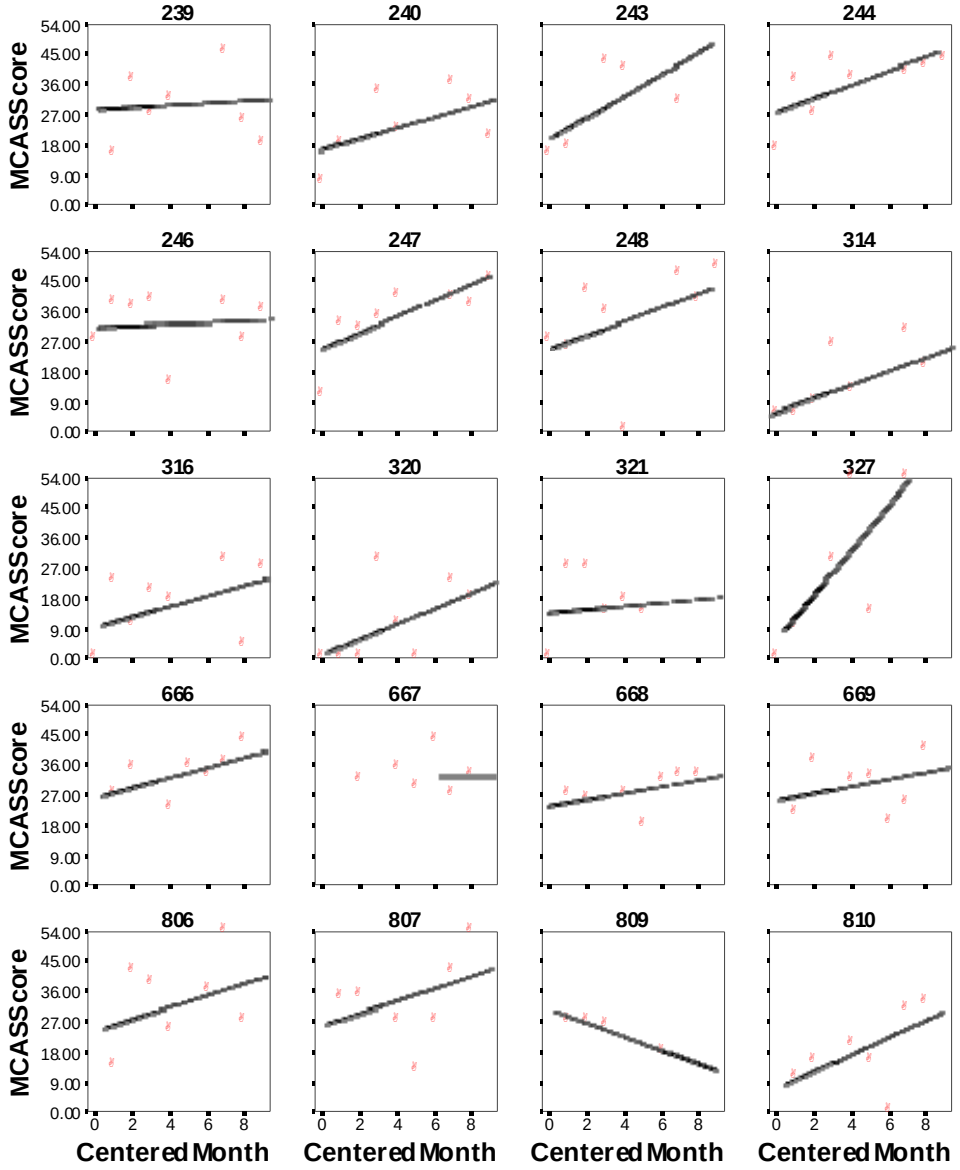


Figure 4.3. Fitting a regression line separately for individual student

Level-1 sub-model:

$$Y_{ij} = \pi_{0i} + \varepsilon_{ij}$$

Level-2 sub-model:

$$\pi_{0i} = \gamma_{00} + \varsigma_{0i}$$

Level-1 sub-model:

$$Y_{ij} = \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij}$$

Level-2 sub-model:

$$\pi_{0i} = \gamma_{00} + \varsigma_{0i}$$

$$\pi_{1i} = \gamma_{10} + \varsigma_{1i}$$

Figure 4.4. Unconditional model (left) and unconditional growth model (right)

4.3.4 Results and discussion

I start by building an "unconditional means model" (left of Figure 4.4) with no predictors, which had a BIC of 31712, as shown in Figure 4.5. This model is called "unconditional" because there are no predictors so that it won't describe change in the outcome over time. Instead, it "simply describes and partitions the outcome 'variation'." This model serves as a baseline by which to evaluate subsequent models. The unconditional means model showed us that the estimated overall average on initial PredictedScore is around 24 points and the estimated variance of intercepts and the estimated variance of residual are both large and statistically significant ($p < 0.05$). This can be interpreted as students' PredictedScore varies over time and students differ from each other on their performance, and there is sufficient variation at both levels to warrant further analysis.

The unconditional model is then compared with an "unconditional growth model" (right of Figure 4.4), in which "CenteredMonth" is introduced as a

predictor TIME. This new model double circled in Figure 4.5 predicts that the estimate of students' average PredictedScore in October, 2004 is about 18 and the average slope is 1.29 ($p < 0.05$), which means after each month, the average PredictedScore increases by 1.29 points, and the increase is statistically reliable. To get more sense out of this number (1.29), I compare this predicted average monthly increase with the improvement required in MCAS test. In MCAS test, students' raw score is out of 54 and their performance is evaluated at 4 levels: warning, need improvement, proficient and advanced. To "jump" from warning to proficient, which is the aim of most students, raw score needs to increase from 23 to 35 and the difference 12 is about 22% of the full score. Distributing the difference across 9 months gives monthly improvement of about 1.33 points, which is slightly higher than the monthly increase predicted in our analysis. The growth model is statistically better having a BIC that is lower by 84 BIC points, which lead us to conclude that *time* is an important factor and also give a positive answer to the first research question.

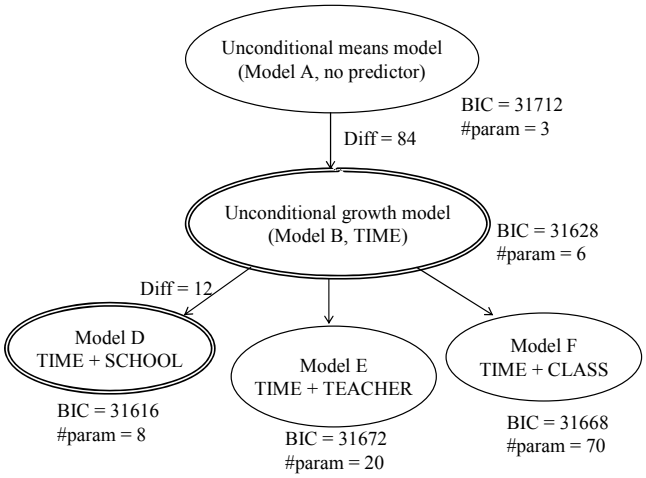


Figure 4.5. Series of longitudinal models for tracking learning

The covariance parameters of the model showed that initial knowledge varies significantly between students after controlling for time ($p < 0.05$) and the knowledge changing rate varies significantly between students. Since the level-2 (i.e. *between-person*) variance components quantify the amount of unpredicted variation in the individual growth parameters (i.e. initial knowledge and rate of change), the significance of the variance indicates there could be other factors we can introduce as level-2 predictors to explain the heterogeneity in each parameter.

New factors are then introduced to build three new models, with each containing one of the **school** (as in Model D), **teache**r (as in Model E) or **class** (as in Model F) variables. The schools model shows a statistical significant advantage. For schools there is a clear difference in incoming students' scores, which makes sense with regard to the fact that one school draws students from the more

affluent side of town. The difference in slope parameter of schools is near significance (p=0.09) and suggests that one school causes more learning in students then another. In contrast to model D, neither model E nor model F displays statistically significant improvement over model B. Our speculation is that **class** may be a level too low to investigate difference on learning, noticing that model F contains 70 parameters that will draw a big increment on BIC values. And a similar problem may associate with model E, which contains 20 parameters.

These results show that the ASSISTments system can reliably track the fact that students are learning in their classes over time. This means that the ASSISTments system can be useful to track and evaluate different interventions. According to these results, not all considered factors (School, Teacher and Class) are significant and we see only that the ASSISTments system detected different rates of learning at different schools. During this analysis, we notice the fact that generally speaking, students with higher estimated initial scores showed lower rates of learning (correlation between intercept and slope of student's learning trajectory is equal to -0.81). The preliminary speculation on this fact is that 1) this may be attributed to the "ceiling effect": it is hard for top students to make fast progress; 2) good students were assigned to Algebra class and learning content that won't be tested until 10^{th} grade and won't appear in the Assistment system. Further investigation needs to be done to explain this phenomenon.

Given that it was the first year of the ASSISTments project, new content is created each month, which introduces a potential confounder of item difficulty. It could be that some very hard items were selected to give to students in September, and students were not really learning but were being tested on easier items. In the future, this confound shall be eliminated by sampling items randomly.

Being able to detect differences between the impacts of various factors can help us improve reporting system as well. Most of the reports described in Chapter 3 are for normal teachers. As a supplementary to those reports, based on the results of longitudinal analysis, I have been working on new reports for principals and administrators which will allow them to see which groups of students need most attention on a wider scope across teachers/classes, based on their gender, special education status, if they get free-lunch and if they are underrepresented. Given these reports, users can also compare teachers and see that which schools/teachers produced more learning than others.

4.4 Combining Dynamic Metrics and Longitudinal Modeling to Predict End-of-year Score Better

Providing instructional assistance in the process of assessing students is the key feature of the ASSISTments. The hypothesis is that the ASSISTments can do a better job of assessing student knowledge than practice tests or other on-line testing approaches by using a "dynamic assessment" approach, thus providing a more precise prediction of student performance on the MCAS test. In section 4.2, I show that by introducing the assistance students required as parameters, we were able to construct a better fitted regression model to predict students' performance on MCAS than simply using their performance on original items. Meanwhile, the longitudinal analysis approach has been described and applied to track student learning over time in section 4.3. In this section, I will show a new method of MCAS score prediction by combining these two parts (i.e. regression model fitting plus longitudinal analysis). Specifically, our research question is: *Can we make a more precise prediction of students' performance on the MCAS by using assistance data longitudinally?*

4.4.1 Introducing new longitudinal outcome variables

As mentioned in Section 4.3, to analyze data longitudinally, one important thing to determine is an outcome whose values change systematically over time. Students' percent correct on the original items is treated as an outcome in Section 4.3. To mimic the real MCAS score, we multiply the percent correct by 54 (the full MCAS score), which makes the outcome range change to 0~54. I will refer to this variable as **plain_predicted_score** in this section to emphasize the fact that it is computed directly from students' monthly performance on original items without any correction. In addition, two new variables, referred to as **original_predicted_score** and **assistance_predicted_score**, will be calculated using linear regression models as I do in Section 4.2. The **original_predicted_score** is regressed on original_percent_correct, while **assistance_predicted_score** is predicted by all the online dynamic metrics as described in Section 4.2. All three predicted scores will be used as the outcome variable individually in the longitudinal data analysis and results will be compared. By introducing dynamic metrics when constructing longitudinal outcome variables, the two techniques are combined together in this new method.

The new approach of MCAS score prediction combines assistance students required and the effect of time. It contains the following steps: a) Split data of the students into training and testing sets; b) train regression models on the training

set and obtain the variables entered in the models and their associated coefficients; c) apply regression models to the testing set and calculate the values of the outcome variables for each student for every month; d) longitudinally track student knowledge using different outcome variables; e) predict student MCAS score given the result of step d); f) compare the outcome variables based on the MCAS score prediction result and answer the research question.

4.4.2 Description of the data

To apply this new approach, I use a data set of 494 students who were using the ASSISTment system from September 16, 2004 through May 16, 2005 for an average of 249 minutes and finished an average of 135 items. I also exclude data from the students' first day of using the ASSISTments. We have the MCAS scores of the 494 students, but no limitation is applied on the total number of problems finished. A paper practice test (I will refer to as *pretest)* was administered in September 2004. Students were asked to finish the test in two periods over two days (totally 80 minutes) and scores of this test were shown to be a significant predictor of MCAS scores (Feng, Heffernan, & Koedinger, 2006a). Here the pretest scores are also used as an independent variable to predict MCAS performance.

Among the 494 students, I let SPSS select approximately 50% as training individuals to train up regression models, leaving 244 students in the testing set.

For the training individuals, a file of 250 rows with one row per student is created. Each row includes variables representing a student's associated real MCAS score, the student's pretest scores, and the 13 "online measures" which we think indicate the amount of assistance a student needs to get an item correct. In contrast to the training set, data for the 244 testing individuals are organized in the "person-period" style to facilitate longitudinal analysis. Because a student only worked on the ASSISTments for one period (about 20 to 40 minutes, varies among schools) every time they came to the lab, similar as in Section 4.2, rather than treating visiting days as the metric for time, all data in one month is collapsed and *month* is used as the level of granularity to measure time to achieve more stable learning-over-time data, and again called "CenteredMonth". It runs from 0 to 7. The "person-period" structured dataset contains on average 5 data waves for each student and values of all the online measures for each CenteredMonth are computed as well.

4.4.3 Modeling

For a long time, we have observed that the ASSISTment system was consistently under-predicting student performance due to the following reasons. Firstly, when building the ASSISTments, authors changed the type of many questions from multiple choices to text input questions, which makes the ASSISTments on average harder than the actual MCAS items. Secondly, the

ASSISTment system always allows students to ask for hints, which to some degree prevents students from trying their best to get the solution. Since hint requests were treated as false responses, this feature could impact students' evaluation. Thirdly, students do not take the ASSISTments as seriously as a real, high-stakes test such as the MCAS and finally they may behave differently when working on a computer because they like or dislike computers (Baker et al, 2005). Therefore, I want to take advantage of regression models to adjust the predicted scores.

First, I run stepwise linear regressions to predict students' real MCAS scores using pretest scores plus original_percent_correct, and pretest scores plus all of the online measures respectively. The models, named Original_Regression_Model and Assistance_Regression_Model, respectively, are summarized in Table 4.9.

Table 4.9. Regression models based upon different independent variables

Model	Parameter	Un-std. Coeff.	Std. Coeff.
Original_ Regression_ Model	(Constant)	4.753	
	pretest	.764	.496
	original_percent_correct	27.869	.367
Assistance_ Regression_ Model	(Constant)	26.035	
	pretest	0.64	.415
	percent_correct	24.205	.307
	avg_attempt	-10.56	-.202
	avg_hint_request	-2.283	-.125

Given Table 4.9, I build the following formulas to compute values for the two new variables that represent student knowledge in a certain month. Noticing that using the above formula, **assistance_predicted_score** takes into account student

performance on scaffolding questions together with the amount of assistance, in particular, the number of attempts and hints, students need on average to get an item correct.

original_predicted_score = 4.753 + pretest * 0.764 + original_percent_correct * 27.869

assistance_predicted_score = 26.035 + pretest * 0.64 + percent_correct * 24.205 - avg_attempt * 10.56 - avg_hint_request * 2.283

Now that the outcome variables are calculated, I fit mixed-effect models on the testing data set and continuously track **original_predicted_score** and **assistance_predicted_score** of students over eight months (Sept, 2004 through May, 2005). The modeling is conducted in SPSS. To facilitate the discussion, I will refer to the two longitudinal models as *Original_Mixed_Model* when original_predicted_score is picked as the outcome variable and *Assistance_Mixed_Model* when assistance_predicted_score is used as the outcome variable. Each model give two parameters for any individual student, intercept (representing initial knowledge status in the first month) and slope (denoting learning rate across the 8 months).

4.4.4 Results and discussion

Recall that my research question asks whether a more precise prediction can be achieved using dynamic metrics longitudinally. To answer this question, I

compute the MAD for both of the two models. Naturally, the predicted scores for the last month (i.e. CenteredMonth = 7) is treated as the predicted MCAS score. For the Original_Mixed_Model, MAD is 6.20, with a standard deviation equal to 4.72 while for the Assistance_Mixed_Model the MAD is 5.533 with standard deviation being 4.40. Consequently, we claim that the Assistance_Mixed_Model, by utilizing the dynamic online metrics, helps to improve the correctness of the prediction on MCAS score. The paired t-test that compares absolute residuals of each student indicates the improvement is statistically significant ($p = 0.011$).

In both regression models presented in Table 4.9, pretest is a significant parameter. We wonder how much the tutoring and assistance information can help without pretest because pretest scores are not always available every school year. I replicate the whole process without using pretest. A comparison of evaluation measures to corresponding values above shows that pretest is an important predictor and without it, the precision of prediction degrades; meanwhile, the model involving tutoring and assistance information still exhibits its superiority and the difference in MAD is almost statistically significant ($p = 0.055$).

In the beginning of the section, I present that we achieve a MAD of 5.533 when predicting MCAS score using the Assistance_Mixed_Model, which is about 10.2% of the full score. To see how good the prediction is, we compare this prediction to the prediction reached by 3 other approaches as measured by MAD

scores. Among other things, pretest scores alone could be used for prediction purposes. So we do a simple regression to predict student's real MCAS scores using associated pretest scores and ended up with a MAD of 6.57 that is statistically significantly higher ($p<0.05$) than 5.533. For a second comparison we look at the predictions in the "Grade Book" reports to teachers on our current web site (Shown in Figure 3.1). The prediction was primitive and was simply a linear function of percent correct on original items. For students in the testing data set, this approach gave a MAD equal to 7.47. In yet a third comparison, we can compare it to using the *plain_predicted_score* as an outcome variable in the longitudinal analysis which brought on a MAD of 9.13. Obviously, all three of these comparisons show higher MAD values, thus indicate that they are not as good at predicting MCAS scores. Note that the comparison between pretest-prediction-method and the ASSISTment approach confounds total time during the assessment (80 vs. 249 minutes) in the sense that it took only about 80 minutes to do the paper and pencil pretest. However, we argue that this is a fair comparison, because our schools (6 schools have adopted the system this year) say they are willing to use the ASSISTments often because they think that students are learning during their use of the ASSISTment web site.

4.5 Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test?

In Section 4.4, I report that we have reduced the MAD by combining dynamic metrics and longitudinal data modeling, but can we do better? Should we be dissatisfied unless we can get a MAD of zero? I want to investigate what a reasonable comparison should be. Ideally, I want to see how good one MCAS test is at predicting another MCAS test. We could not hope to do better than that (But considering there is measurement error in the MCAS test, maybe we can!). However, I do not have access to data for a group of kids that took two different versions of the MCAS test to measure this. Then, how can I evaluate the predictive accuracy of a student model relative to the standardized test?

4.5.1 The "splitting" method

In Feng, Heffernan, & Koedinger (2006b), I ran a simulation study. In the study, I took students' scores on 2005 MCAS test, randomly split the test into two halves, and then used their scores on the first half to predict the second half. I call this method the "splitting" method. In the study, open response questions from the MCAS test are excluded, while only the remaining 34 multiple-choice and short answer questions are kept with regard to the fact that open response questions are not supported in the ASSISTments system. Then the 34 items are randomly split

into two halves and student performance on one half is used to predict their performance on the other half. This process was repeated 5 times. On average, I got MAD of 1.89, which is about 11% of the full score (17 points with one point for each item). Hence I drew the conclusion that using the new approach I have established, the prediction of MCAS score is as good as the real MCAS test itself, with the caveat that only 34 items are utilized in the process here, while my prediction models are built based on students' work on 135 ASSISTment items over eight months.

4.5.2 The "longitudinal" method

Later on, I propose another approach to use student cross-year data for determining whether a student model is as good as the standardized test at estimating student proficiency (Feng, Beck, Heffernan, & Koedinger, 2008). I call this method the "longitudinal" method.

Assume student math proficiency in the 8th grade and in the 10th grade are highly correlated. Since the measurement error is relatively independent due to the two years time interval between the tests, therefore, whichever (our student model or the MCAS test) better predicts 10th grade MCAS score is better assessing student math skill at 8th grade.

4.5.2.1 Results

Let's define MCAS8' be the leave-one-out[21] predicted score for 8th grade MCAS that comes from the mixed model in Section 4.2; MCAS8 be the actual 8th grade MCAS score and MCAS10 be the actual 10th grade MCAS score. Then we ask the question: *Can MCAS8' predict MCAS10 better than MCAS8 does?* To answer the question, we calculated the correlation between the three metrics: MCAS8', MCAS8 and MCAS10, as presented in Figure 4.6.

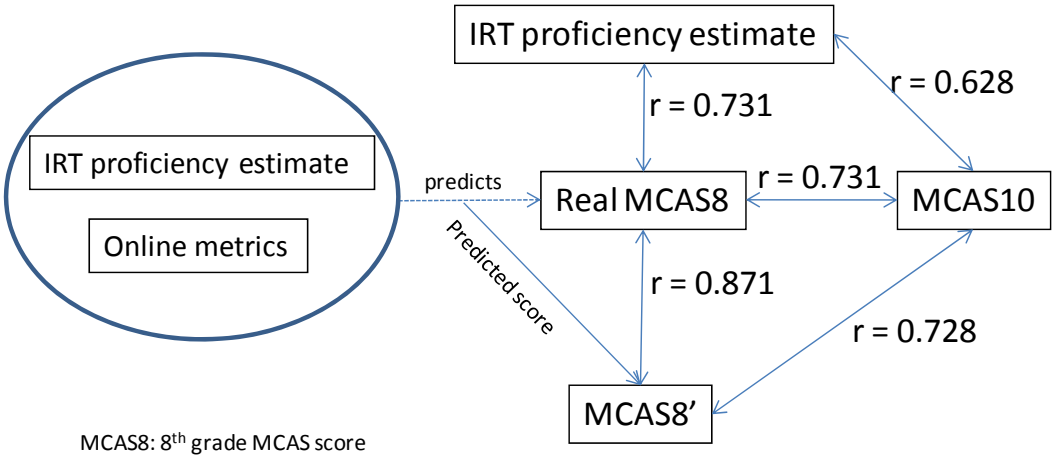


Figure 4.6. Correlation between IRT student proficiency estimate, MCAS8', MCAS8 and MCAS10

First of all, I want to point out that all correlations in Figure 4.6 are statistically reliable ($p < 0.001$). The student proficiency estimated by the lean model (as described in Section 4.2) correlates with MCAS10 with r equal to .628.

[21] The adjusted predicted score is calculated by doing "leave-one-out" cross validation in SPSS.

It does not do as well as MCAS8 and MCAS8' as we have expected. Even though, we think it is worth finding out and having this lean model, which is based on less data, as a contrast case. It is the most direct test of the question of whether ASSISTment use could essentially replace the 8th grade test. Both MCAS8 and MCAS8' are reliable predictors of MCAS10. MCAS8 correlates with MCAS10 with r equal to 0.731 while the correlation between MCAS8' and MCAS10 is fractionally lower (r = 0.728). A significance test[22] shows they are not statistically reliably different, which suggests that our student model can do as well as MCAS test on predicting the MCAS score two years later. Since both MCAS tests are measuring the student's math proficiency, it can be considered as the evidence that the student model is doing a good job estimating student math proficiency. At the very least, what our system is modeling is relatively stable across a two-year interval.

4.5.3 The "proxy" measure method

A third approach of comparing a student model to a standardized test has been described in the literature. Beck & Sison (2006) found 3 tests that measures extremely similar constructs to the standardized test that they were interested in. They took the arithmetic mean of those tests as a proxy measure for the true score

[22] The test is done online at http://www.quantitativeskills.com/sisa/statistics/correl.htm

on the original measure. I will refer to this method as the "proxy" measure method.

4.5.4 Discussion

Put the three methods that have been described above together, we can see that the pro of the "proxy" measure method is that it can be done quickly while the con is that construct validity could be an issue. Similarly to the "proxy" measure method, the pro of the simulated "splitting" method is the quickness but it also has some cons. Firstly, if there is measurement error for a particular day (e.g. a student is somewhat ill or just tired), then splitting the test in half will produce a correlated measurement error in both halves, artificially increasing the test's reliability relative to the measure of the "longitudinal" method (which is not based on data from the same day as the MCAS). Secondly, to do the splitting, it requires assess to item level data which is not always available. The "longitudinal" method, by going across years, avoids this confound with measurement error, and get a fairer baseline. Though, we do admit that it takes longer time and harder effort to collect data across years (in our case, 3 years). I want to stress that this is a rather long-term prediction. The collection of the online data started in September, 2004; the 8^{th} grade MCAS score that we are predicting came in at the end of year 2005; while the 10^{th} grade MCAS score that we used to evaluate our prediction were available at the end of year 2007.

In conclusion, I have proposed new methods for evaluating the predictive accuracy of a student model relative to the standardized test, using a simulation study or using student standardized test scores across multiple years. And I argue that I can model student math proficiency as well as the standardized test as measured by the new evaluation criteria.

4.6 Related Work and Contribution

An intelligent learning environment adapts the educational interaction to the specific needs of the individual student. Thus, student modeling is an essential component in such an environment and the learning effectiveness depends heavily on the understanding of student knowledge, difficulties, and misconceptions. Yet, assessing students automatically, continuously and accurately without interfering with student learning is an appealing but also a challenging task.

Even before the computer supported systems become popular, much work has been on developing "testing metrics" for dynamic testing (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2001, 2002) to supplement accuracy data (wrong/right scores) from a single sitting. Researchers have been interested in trying to get more assessment value by comparing traditional assessment (static testing; students getting an item marked wrong or even getting partial credit) with a measure that shows how much help they needed. Bryant, Brown and Campione

(1983) compared traditional testing paradigms against a dynamic testing paradigm. Grigorenko and Sternberg (1998) reviewed relevant literature on the topic and expressed enthusiasm for the idea. In the dynamic testing paradigm, a student would be presented with an item and when the student appeared to not be making progress, would be given a prewritten hint. If the student was still not making progress, another prewritten hint was presented and the process was repeated. Sternberg & Grigorenko (2001, 2002) argued that dynamic tests not only serve to enhance students' learning of cognitive skills, but also provide more accurate measures of ability to learning than traditional static tests. In Bryant, Brown and Campione's study they wanted to predict learning gains between pretest and posttest. They found that student learning gains were not as well correlated (R = 0.45) with static ability score as with their "dynamic testing" (R = 0.60) score. They also suggested that this dynamic method could be effectively done by computer, but never pushed toward to conduct such studies in a computer system. The ASSISTment system provides an ideal test bed to investigate dynamic assessment as it automatically provides students with feedback, scaffolding questions and hints. So, naturally, I extend these prior works and test the idea of dynamic assessment in ASSISTments. As far as I know, only very few work (e.g. Beck, Jia & Mostow, 2006) has noticed the nature and amount of help that a student is given appear to be of value in judging a student's mastery of knowledge, and test the idea in a real computer system. Thus, I claim that operationalising the

idea and assessing whether this promise is achievable in an implemented system is an important contribution.

Beck, Jia & Mostow (2004) were able to increase the within-grade correlation between their model and student performance on a fluency test significantly by adding help-seeking behavior in the computer Reading Tutor (Mostow & Aist, 2001). In psychometrics field, van der Linden and his colleagues (Klein Entink, Fox, & van der Linden, in press; van der Linden, 2007, 2008) have extended the traditional IRT model to take into account student response time. My work is different to these works in that I take more features into account in the same model. An important contribution is that I contrast models based upon correctness information with models based only upon dynamic metrics, and show the later is superior at assessing student knowledge, which has been totally ignored in traditional assessment practice.

Anozie & Junker (2006) pursued a rather different approach from me, looking at the changing influence of online dynamic metrics on MCAS performance over time. They computed monthly summaries of online metrics similar to those used by us, and built several linear prediction models, predicting end-of-year raw MCAS scores for each month.

Modeling student response data from intelligent tutoring systems has a long history (e.g. Corbett, Anderson, & O'Brien, 1995; Draney, Pirolli, & Wilson,

1995). Below, I give a quick literature review of works on student modeling. These works typically used different approach, such as knowledge tracing, Bayesian networks, or have been applied in different learning architecture, such as distributed learning environments, from what I have applied in this dissertation.

One stream of modern educational research is on the modeling of student's individual problem-solving performance. Such analysis provides detailed assessments of student competence at different skills, and the results are usually used to guide the selection of next instructional actions in ITS. Corbett & Anderson (1995) proposed a process called *knowledge tracing* to model students' changing knowledge state during skill acquisition using a two-state Markov model. They showed the model was very successful in predicting test performance. Corbett, Anderson, & O'Brien (1995) further explored the quality of student modeling in the ACT programming tutor. Corbett, Anderson & O'Brien (1995) observed the power function might not hold for some complex skills thus there are blips in the learning curves. On top of that they found complex skills can be decomposed into sub-skills which result in smoother learning curve. This work led some cognitive scientists to look back at existing cognitive model and to bring up methods to improve cognitive models (e.g. learning factor analysis, Cen, Koedinger & Junker, 2005, 2006).

Bayesian networks are becoming an increasingly popular way of representing the state of a student's knowledge, skills, or abilities, especially in intelligent learning environments. Mislevy and colleagues (e.g. Mislevy, 1995; Mislevy & Gitomer, 1996; Mislevy, Steinberg, & Almond, 2003) focused on the role of probabilistic reasoning in ITS. They suggested evaluating students and providing feedbacks to students and teachers based on evidence. Conati & VanLehn (1997) used Bayesian networks to model student knowledge and updated the networks in real time. In addition to skill assessment and performance prediction, the model was also used to do plan recognition, and was used by the Help component of ANDES system (VanLehn, 2005) to provide tailored support. The Bayesian network for a problem was automatically constructed from the solution graph produced by a problem solver. The display capability of Bayesian networks is designed to work with individual student one at a time. Yet, Almond et al. (2008) viewed the problem from a teacher's perspective, concerning how to make inferences about a group of students. To ease the complicated coding effort involved in the usage of Bayes nets, esp. dynamic Bayes nets, Chang et al. (2006) introduced a Bayes Net Toolkit for Student Modeling (BNT-SM) that allows a researcher to describe causal relationships among student knowledge and observed behavior. Later on in Beck et al. (2008) dynamic Bayesian networks was trained up using the toolkit for evaluating the efficacy of assistance that was provided to students in an ITS that tutors reading.

Student modeling (or user modeling) is a distinctive feature of user-adaptive software systems, including adaptive hypermedia and other adaptive Web systems, and has caught attention of many researchers. For instance, Brusilovsky, Sosnovsky, and Shcherbinina (2005) addressed the issue of user modeling in a distributed learning architecture. They described a generic student modeling server and introduced a specific, topic-based knowledge modeling approach that was used in an adaptive system to help students to select the most relevant self-assessing quizzes. Brusilovsky and Millán (2007) focused on user modeling in adaptive systems. They explored the nature of the information being modeled in adaptive web (including users' knowledge, interests, goals and tasks, background, individual traits, and context of work), the overlay modeling approach, and also uncertainty-based user modeling for adaptive systems.

Another stream of student modeling in learning environment has to do with the study of open learner models (Bull, 2004; Bull & Kay, 2007). Student models (or learner models) are usually not accessible to the students themselves. However, some works integrated the educational benefits of allowing students to access the learner model contents, and even to negotiate with the system on the understanding of their knowledge proficiency. It was argued that this exposure will increase students' self-awareness of their knowledge and, thus, enhance learning (e.g. Kay, 1997; Bull & Nghiem, 2002). Certain amount of work has been done on application and evaluation of open learner models in various

learning environments (e.g. Dimitrova, Self, & Brna, 2001; Mitrovic & Martin, 2002; Bull, & McEvoy, 2003; Bull, et al., in press; Wongchokprasitti & Brusilovsky, 2007; Ahn et al. 2007).

Among these studies on modeling student proficiency, different criteria have been used to evaluate the effectiveness of the student models. For instance, Beck & Sison (2006) used relative closeness to real scores and mean absolute error (MAE); Anozie & Junker (2006) evaluated their model using mean absolute deviation (MAD) which is essentially the same as MAE; while we have reported R square and Bayesian Information Criterion (BIC) (Feng, Heffernan & Koedinger, 2009).

4.7 Conclusion

In this chapter, testing challenge in the ASSISTments system, a web-based assessment and learning system is addressed. I bring up a few research questions focusing on the assessment ability of the system. The questions are answered by mining our log data and comparing with standardized test results. Solid evidence is presented that the online assessment system does a better job of predicting student knowledge than traditional approaches by being able to take into consideration how much tutoring assistance was needed, how fast a student solved a problem and how many attempts were needed to finish a problem. And the

prediction accuracy is comparable with results from the standardized test. Promising evidence is also found that the online system is able to track students' learning during a year well.

Chapter 5

5 Cognitive Diagnostic Assessment and Skill Model Analysis

Chapter organization: This chapter is organized as follows. Section 1 motivates the work of building skill models from the perspective of both practitioners and builders of intelligent tutoring systems. The second section describes how we have constructed a fine grained model and built a hierarchal structure of models at different granularities. Section 3 analyzes the relative effectiveness of skill models at various grained level and shows evidence that the fine grained model leads to more accurate prediction of external state test scores. Section 4 compares my work to other related works in the literature and point out the contribution of my works in this chapter. Section 5 concludes the chapter. Contents of this chapter draw largely from the following papers: Feng, Heffernan, Mani, & Heffernan (2006), a paper presented at AAAI'06 workshop, Razzaq, Feng, Pardos, & Heffernan (2007), the article published in Journal of Technology,

Instruction, Cognition, and Learning, and Feng, Heffernan, Heffernan, & Mani (2009), an article recently published in IEEE Transaction on Learning Technologies.

5.1 Introduction

Most large standardized tests are "unidimensional" in that they are analyzed as if all the questions are tapping a single underlying skill. However, cognitive scientists such as Anderson & Lebiere (1998) believe that students are learning individual skills. Among the reasons that psychometricians analyze large scale tests in a unidimensional manner is that students' performance on different skills are usually highly correlated, even if there is no necessary prerequisite relationship between these skills. Another reason is that students usually do a small number of items in a given setting (for instance, 39 items for the 8th grade math MCAS test), which makes it hard to acquire identifiability for each single skill, especially when the number of skills that need to be mastered is larger than the number of the items in the test. Such tests work pretty well at telling you which students are performing well but if we want to give help to students, we need a detailed understanding of the concepts and skills involved in the problems students are trying to solve. As mentioned in Chapter 3, one of the issues with the report shown in Figure 3.1 is the *uninformative* issue. Teachers will be happy to

know that we can accurately predict student's total score in MCAS. But, on top of that, they will also want to be given more specific information on how to adjust their instruction to help those students who are predicted to fail the test. Not only do they want to know these students are having difficulties, but also they need to be told where exactly the difficulties lie so that they can isolate and try to eliminate them in their instructional practice.

A task analysis of the problems shown in Figure 1.2 is that it requires the following knowledge components: Know the concept of congruence; know how to parse an algebra expression and solve equations; know the concept of perimeter; know how to add and subtract numbers. These knowledge components make up a skill model for this problem. It is easier to list these skills (or concepts) than it is to determine which are the hardest for students to learn. However, understanding the hard skills for students is important as such skills should be the focus of assessment and instruction. Assessment is more accurate if problems are tagged with the hardest skills required because failure is more likely a consequence of a lack of that skill than any others. Meanwhile, instruction becomes more effective when focusing on supporting learning of such skills, because these skills need the most attention and benefit the most from designing scaffolds that support their acquisition. Yet, literature in the area of difficulty factors assessment suggests that how identifying what is difficult for students is not at all obvious (e.g. Heffernan & Koedinger, 1997, 1998). By developing scaffolds and tagging all questions in

fine grained skill models in an automated assessment environment as ASSISTments, we hope to address the hypotheses about the nature of student learning difficulties, and meanwhile, report more informatively to classroom instructions.

From the perspective of the needs of intelligent tutoring systems, one key component of creating an intelligent tutoring system is forming a model that monitors student behavior. ITS needs the construction of complex models to represent the skills that students are using and their knowledge states. As students work through the program, the model tracks their progress and chooses what problems will be displayed next. By using a better skill model, a system should be able to do a better job of predicting which items students will get correct in real-time. That means the system can do a better job of selecting the next best item for students to work on. For instance, one criterion of the next "best" item could be the one that has the largest ratio of expected test-score gain to expected time to complete the problem. Expected test score gain will be a function that depends upon both the expected rise in skills from doing that item at that time, as well as the weight of those skills on the test (i.e., the MCAS). A better model would also help to address the issues as we mentioned above to help teachers adjust their instruction in a data-driven manner. Such a model will allow a teacher who has one week before the MCAS to know what topics to review to maximize the class average. We can make a calculation averaging the whole class to suggest what

will give the teacher the biggest "bang for the buck." An example of a useful report (Feng & Heffernan, 2007) that teachers can get using the ASSISTments system is shown in Figure 3.5. Teachers can see how their students are doing on each skill and can determine where they need to spend the most time. To summarize, a fine-grained model that is based upon a thorough understanding of what is hard for students can lead a better categorization of test items and better guidance for teaching.

I am engaged in an effort to investigate if we can do a better job of predicting the state test scores by modeling individual skills in a finer grain size (Feng, Heffernan, Mani, & Heffernan, 2006; Feng, Heffernan, Heffernan, & Mani, 2009). This is not applicable in the traditional assessment environment because of limited testing time and test items. As a consequence, it is hard to determine which skill(s) to credit or blame, especially when a wrong answer is given. However, the special structure of the ASSISTment System gives teachers that very information. By providing scaffolding questions within test items, we can independently assess individual components of knowledge and not just overall performance on an item. Since original questions are always followed by scaffolding questions, each addressing a single piece of the knowledge, when a student answered the original question wrong, we can rely on his responses to scaffolding questions to make precise inferences about exactly where the student has a misunderstanding and are able to track the specific knowledge component

precisely. Moreover, students are using the ASSISTments System regularly in their normal math class during a school year, working on mathematic questions drawn from a pool of more than 1,400 questions generated from released state test items, local teachers, and project staff. The continuous usage allows us to collect more evidence of student's performance on every skill. I hypothesize that richer information per item, accumulated over many sessions, will compensate for the smaller number of MCAS-like items that students are exposed to at any one sitting.

5.2 Development of a Fine Grained Skill Model

Cognitive modeling involves a great deal of detailed protocol collection and task analysis. The models are not easy to construct and are difficult to verify. Yet, cognitive modeling is very important in an ITS as it is the basis of cognitive diagnostic assessment and teacher reporting. Our initial effort on building fine grained models is based upon our knowledge of the existing background research from mathematics education as well as cognitive task analysis conducted by subject-matter experts. In April, 2005, we staged a seven hour long "coding session," where we invited our subject-matter experts to create a set of skills and to use those skills to tag all of the existing 8th grade MCAS items. Because we

wanted to be able to track learning between items, we wanted to come up with a number of skills that were somewhat fine-grained but not too fine-grained such that each item had a different skill. We therefore imposed the limit that no one item would be tagged with more than three skills[23]. It is not coincidence that many of the ASSISTments System questions have three scaffolding questions; we wanted the fine grained-ness of the modeling to match the fine grained-ness of the scaffolding. We wanted the scaffolds to have identifiability, meaning that each scaffolding question should be tagged with only one skill. Identifiability is important because when a student got a question tagged with two skills wrong, we will have a hard time coming up with a method that would be able to blame the lack of a single skill. As a matter of fact, in the data set to be used in Section 5.3, the average number of skills tagged to main questions is 1.44 and the number is 1.03 for scaffolding questions, which means that for many questions in ASSISTments System, there is only one skill tagged with the main question. In such cases, each of the scaffolds was also tagged with that skill. There clearly is something a bit odd there, as an individual scaffolding question should be easier than the main item. I noticed this, but my modeling effort does not correct for the presumably wrong assumption that questions tagged with the same set of skills have the same degree of difficulty.

During the "coding session." the subject-matter expert was free to create whatever skills she thought appropriate. She reviewed the items, solved the problems and, thus, conducted a cognitive task analysis to identify what knowledge was required to perform each task. When the coding session was over, we wound up with about a model of 106 skills. Yet since only 78 out of the 106 skills were involved in the data used in this dissertation so I call this model the WPI-78. To create the coarse-grained models, the fine-grained model was used to guide us. We decided to use the same broad strands that are used by both the National Council of Teachers of Mathematics and the Massachusetts Department of Education. These strands are named 1) "Patterns, Relations and Algebra," 2) "Geometry," 3) "Data Analysis, Statistics and Probability," 4) "Number Sense and Operations" and 5) "Measurement." The Massachusetts Department of Education actually tags each item with exactly one of the five strands, but our mapping was inferred from the WPI-78, so it was not the same as the state's mapping. It hereby was named WPI-5. Furthermore, we allowed multi-mapping, i.e., allowing an item to be tagged with more than one skill. An interesting piece of future work would be to compare our fit with the classification that the state uses. Similarly, we adopted the name of the 39 learning standards (nested inside the five strands) in the Massachusetts Curriculum Framework, associated each skill in WPI-78 to one of the learning standards, thus we created the model WPI-39. After the students had taken the state tests, the state released the items in that

test, and we had our subject-matter expert tag up these items in WPI-5, WPI-39 and WPI-78.

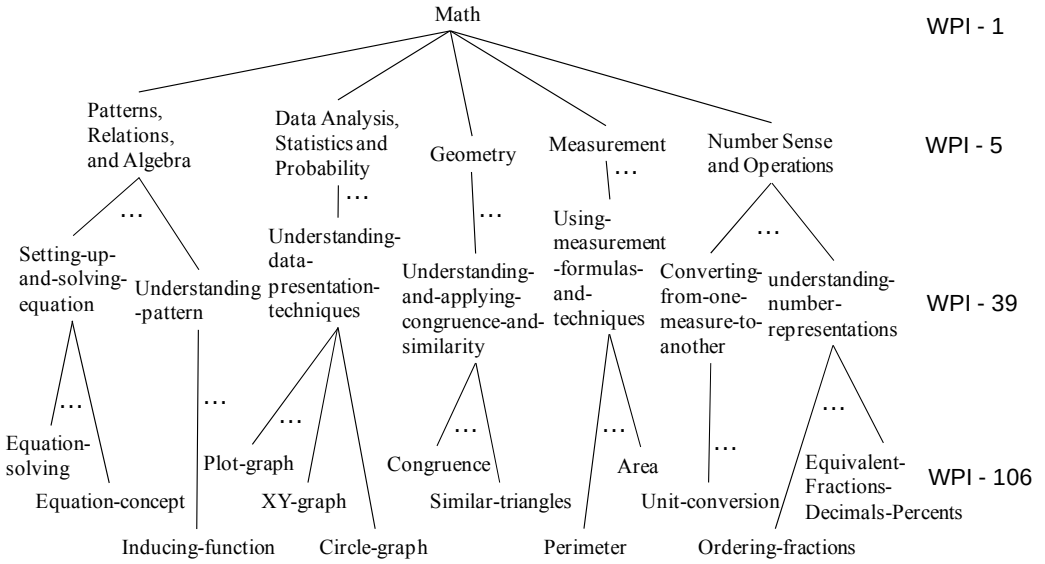


Figure 5.1. The hierarchal structure of four skill models at different granularities

As mentioned above, the WPI-1, WPI-5 and WPI-39 models are derived from the WPI-78 model by nesting a group of fine-grained skills into a single category. Figure 5.1 shows the hierarchical nature of the relationship among WPI-78, WPI-39, WPI-5, and WPI-1. The bottom layer lists 13 out of the 78 skills in the WPI-78 model. For instance, both nested inside the Setting-up-and-solving-equation skill in the WPI-39 model, *equation-solving* is associated with problems involving setting up an equation and solving it; while *equation-concept* is related to problems that have to do with equations in which students do not actually have to solve them. In the upper layer we see how the two skills in WPI-78 are nested

inside of “Patterns, Relations and Algebra,” which itself is one piece of the five skills that comprise the WPI-5 skill model.

As shown in Figure 2.1, in the WPI-78 skill model, the first scaffolding question of item 19 from the 2003 MCAS test is tagged with “congruence,” the second tagged with “perimeter,” the third tagged with “equation-solving.” The original question is associated with all three skills. When a student answers the original question wrong, we will examine his response to the scaffoldings to determine which skill is causing difficulty. According to Figure 5.1, in the WPI-39 model, the corresponding skills tagged are “Setting-up-and-solving-equations,” “Understanding-and-applying-congruence-and-similarity,” and “Using-measurement-formulas-and-techniques.” In the WPI-5, the questions are tagged correspondingly with “Geometry,” “Measurement” and “Patterns, Relations and Algebra,” and just one skill of “math” at the WPI-1.

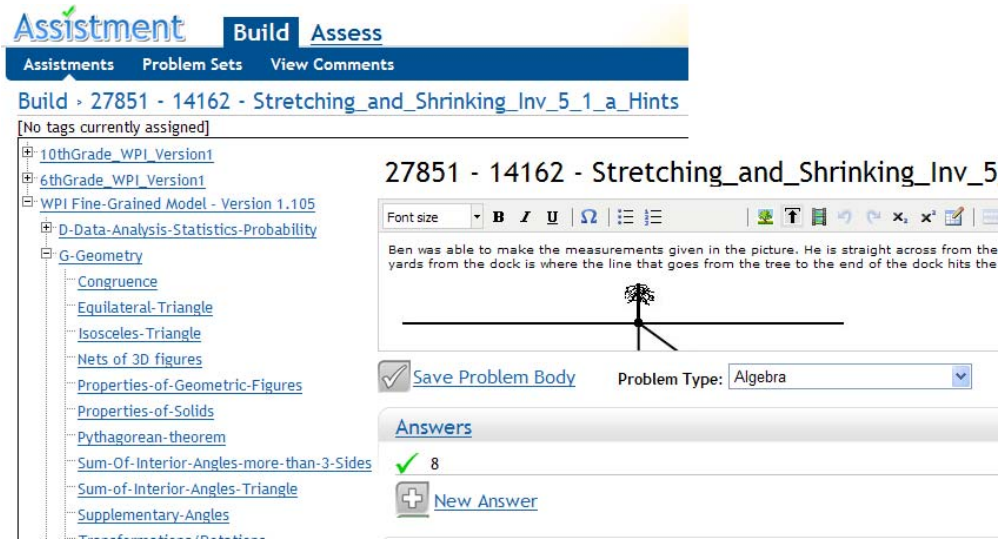


Figure 5.2. A screen shot showing an item and a list of skills that could be tagged to the item in the builder

The ASSISTment Builder (Heffernan et al., 2006) provides technology support for authors to tag skills for the ASSISTment System question they build. This tool, shown in Figure 5.2, provides a means to link certain skills to problems and to specify that solving the problem requires knowledge on that skill. The skills are organized in a hierarchical structure. The authors are allowed to browse the skills within each model and to map the ones they select to a problem.

5.3 Do We Need Fine Grained Models?

Many intelligent tutoring systems use fine-grained models to track student mastery of individual knowledge components. For instance, the Cognitive Tutors

(Ritter, et al., 2009) tracks 2,400 skills in four curricula. Do we really need such kind of fine-grained models? Barnes has done considerable work with trying to induce transfer models, in this work called q-matrices from data (Barnes, 2005, 2006). Winters et al. (2005) has compared a variety of statistical approaches for constructing transfer models. They both found sparse models generated by computers. However, both of their approaches ignored learning, which seems to be a big part of skill mapping in cognitive models. McCalla & Greer (1994) pointed out that the ability to represent and reason about knowledge at various levels of detail is important for robust tutoring. Gierl, Wang & Zhou (2008) proposed two directions for future research of cognitive assessment, and one is to increase understanding of how to specify an appropriate grain size or level of analysis with a cognitive diagnostic assessment (Nichols, Chipman & Brennan, 1995; Leighton, & Gierl, 2007). In this section, I consider four skill models with different granularity, including a unidimensional model and a fine-grained model developed at WPI with 78 skills. The four models are structured with an increasing degree of specificity as the number of skills increases. The measure of model performance is the accuracy of the predicted MCAS test scores based on the assessed skills. Given that the fine-grained model is composed of 78 skills, people might think the model would naturally fit the data better than the skill models that contain far less skills, maybe even overfit the data with so many free parameters. However, I evaluate the effectiveness of the skill models over a

totally different data from MCAS tests, the external state tests as the testing set. Predicting students' scores on this test will be our gauge of model performance. Hence, I argue that overfitting would not be a problem in our approach.

5.3.1 Description of the data

For the 2004-2005 school year, I collected data from 447[24] students, who used our system from Sep. 17, 2004 through May 16, 2005 for on average 7.3 days (one period per day)[25]. All these students worked on the system for at least six days (one session per day). Again, data from the students' first day of using the system is excluded. The item-level state test report is available for all these 447 students so that I am able to construct our predictive models on these students' data and evaluate the accuracy on state test score prediction. The original data set, corresponding to students' raw performance (before applying any "credit-and-blame" strategies as described in the next section and not inflated due to the encoding used for different skill models), contains about 138 thousand data points, among which around 43 thousand come from original questions. On average, each student answered 87 MCAS (original) questions. I will refer to this

24 The amount of data is limited by the maximum memory allowed by the open source statistical package R. I reduce the number of students from the initial 495 to 447 to accommodate the modeling process in R. The 447 students are randomly selected from the whole data set.

25 Given the fact that the state test was given on May 17, 2005, it would be inappropriate to use data after that day for the purpose of predicting state scores. Therefore that data was not included in our data set.

data set as DATA-2005. For the students involved in DATA-2005, the mean score out of 34 points is 17.9 (standard deviation=7.1). 7

Similarly, I build a data set for the usage of the ASSISTment System during the 2005-2006 as well. The data set involves 474 students who, on average, worked in the system for 5.5 days and answered 51 original questions. The item level response data from the 2006 MCAS tests is available for these students as well. This data set will be referred to as DATA-2006. For the students in 2006 data set, the mean score is 18.8 (standard deviation = 7.8).

Both of the data sets are organized so there can be one or multiple rows for every student response to each single question depending on which skill model we are interested in and how many skills the question is "tagged" with in that particular skill model. For instance, suppose a question is tagged with two skills in a model, then for each response made to the question there would be two rows in the data set, with skill names listed in a separate column. Students' exact answers are not included. Instead, we use a binary column to represent whether the student answered the specified item correctly or not. No matter what the input type of the item is (multiple-choice or text-input), a "1" indicates a correct response while a "0" means a wrong one. Additionally, a column is associated with each response, indicating the number of months elapsed since September 17, 2004 (or September 1st, 2005) till the time when the response was made. Thus in

DATA-2005, the number of months elapsed for a response made on September 17th will be zero, and the number will be 1 for a response made at October 17th, 2004, and so on. This gives us a longitudinal, binary response data set across the school year.

Table 5.1. Sample raw data of one student

Row ID	Student ID	State Test ID	question ID	WPI-78 skills	Original?	Response	Month Elapsed
1	950	2003- item 19	326	Congruence	Yes	0	1.32
2	950	2003- item 19	326	Perimeter	Yes	0	1.32
3	950	2003- item 19	326	Equation-Solving	Yes	0	1.32
4	950	2003- item 19	327	Congruence	No	0	1.32
5	950	2003- item 19	328	Perimeter	No	1	1.32
6	950	2003- item 19	329	Equation-Solving	No	0	1.32
7	950	2003- item 19	330	Equation-Solving	No	0	1.32
8	950	1999- item 27	1183	Perimeter	Yes	0	2.94
9	950	1999- item 27	1183	Area	Yes	0	2.94
10	950	1999- item 27	1184	Perimeter	No	1	2.94
11	950	1999- item 27	1185	Area	No	1	2.94

Table 5.1 displays 11 rows of the raw data for one student (system student ID = 950) who finished the item 19 (from 2003 state test, shown in Figure 1) and item 27 (from 1999 state test) on two different days. The first 7 rows represent the student' responses to item 19 (with system ID for the original question being 326) and the remaining 4 rows show his responses to item 27 (with system ID for the original question being 1183). We can see that since the original question of item 19 was tagged with 3 skills, "Congruence," "Perimeter" and "Equation-Solving," the student's response was duplicated in rows 1 – 3. Likewise, the original question of item 27 is tagged with 2 skills as shown in row 8 and row 9. If a student gives correct answer to an original question, the response column of all rows corresponding to that original question is marked as "1". If the student

answered the original question wrong, he gets "0" in all the corresponding rows, and we will use his response to separate scaffolding questions to determine which skill to blame (assuming each scaffolding question is tagged with only one skill). In the example, the student answered both the original questions incorrectly, thus got "0" in the response column of rows 1-3 and rows 8-9), and was presented with the scaffolding questions. The student did not do very well on the first item. He only gave a correct answer to the second scaffolding question (indicated by "1" in the response column of row 5), and failed on all others. In contrast, although the student did not get item 27 right on the first shot on the original question, the student wound up answering both scaffolding questions correctly. The WPI-78 is the skill model used here.

5.3.2 Mixed-effects logistic regression modeling

The mixed-effects logistic regression model is fit, a longitudinal model, on the data to obtain an estimate of student knowledge on individual skills at a certain time, and for simplification, assuming student knowledge changes linearly over time. The mixed-effects logistic regression model is also referred to as the Generalized Linear Mixed-effects Model (GLMM) in the statistics literature and as hierarchical models in Bayesian settings. It consists of both *fixed effects*, parameters corresponding to an entire population or repeatable levels of factors, and *random effects*, parameters corresponding to individual subject drawn randomly from a population. For dichotomous (binary in our case) response data, several approaches have been developed. These approaches use either a logistic

regression model or a probit regression model and various methods for incorporating and estimating the influence of the random effects on individuals. Since we want to track individual student's development of skills over time and make predictions, we chose mixed-effects logistic regression model because it takes into account the fact that responses of one student to multiple items are correlated, moreover, the random effects allow the model to learn parameters for individual students separately. Hedeker & Gibbons (2006) described mixed-effects models for binary data that accommodate multiple random effects. As these sources indicate, the mixed-effects logistic regression model is a very popular and widely accepted choice for analysis of dichotomous data.

As a statistical foundation of the mixed-effects generalization of the logistic regression model, we first present the simpler fixed-effects logistic regression model. Let p_i represent the probability of a positive response on an item for the ith individual. The probability of a negative outcome is then $1 - p_i$. Let $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the set of covariates and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ be the vector of corresponding regression coefficients. Then the logistic regression model can be written as:

$$p_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

The model can also be represented in terms of log odds or *logit* of the probabilities, namely:

$$\log[\frac{p_i}{1-p_i}] = x_i^{'}\beta$$

In logistic regressions, the *logit* is called the link function because it maps the (0, 1) range of probabilities onto (-∞, +∞) range of linear predictors. And by doing this, now the logistic regression model is linear in terms of the logit, though not in terms of the probabilities.

The simple logistic regression model can be generalized to be the mixed-effects model by introducing the random effects. Suppose TIME is the only covariate we care about in the model (*skill* can be introduced as a factor in the model in a similar way). The 2-level representation of the model in terms of *logit* can be written as

Level-1 (or within-person) model:

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

Level-2 (or between-person) model:

$$b_{0i} = \beta_0 + v_{0i}$$
$$b_{1i} = \beta_1 + v_{1i}$$

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

$TIME_{ij}$ refers the *j*th opportunity when student *i* answered a question. In our data, it is a continuous value representing the number of months (assuming 30 days in a month) elapsed since student *i* took his first action in the system

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

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

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

Such a model is often referred to as a "longitudinal model" (Singer & Willett, 2003; also see Chapter 4 for an introduction of longitudinal data analysis) since TIME is introduced as a predictor of the response variable, which allows us to investigate change over time. The models are fitted in R (R Development Core Team, 2007) using *lmer()* function in *lme4* package (Bates, 2007) and "logit" is used as the link function. I introduce skills as fixed-effect factor and TIME (*monthElapsed*) as both a fixed effect and a random effect in order to learn both

the learning rate per month for the whole group of students on average and the variation of each individual student. I also include the interaction between *skills* and *monthElapsed*, which tells the model to learn students' average learning rate separately for each skill. Notice that skills are not included as random effects, which meant the model assumes a student's learning rate did not vary over different skills[26].

After the model is constructed, the fixed-effects for the whole group (i.e. β_0, β_1 in the above 2-level model), including an intercept, a coefficient for the *monthElapsed* covariate, four coefficients for the *skills*, one for each skill in the WPI-5 model, and four coefficients for the interaction term, and the random effects for each student (i.e. v_{0i}, v_{1i}), including an intercept indicating a student's incoming knowledge and a slope (coefficient for *monthElapsed* as a random effect) indicating the student's overall learning rate per month, are extracted. Then the two learning parameters "intercept" and "slope" (i.e. b_{0i} and b_{1i} in the model above) are calculated for each individual student and for each skill. Given this, I can apply the model on the items in the state test to estimate students' response to each of them.

[26] This is just a simplifying assumption. Of course, in reality, it is possible that a student might learn one skill (e.g. perimeter) faster than another one (e.g. congruence).

5.3.3 Predicting state test scores

After the model is fit, we have skill levels of the students based on their online ASSISTment data using the specified skill model. I then apply the model on the actual state test. All the items in the state tests have been tagged in all of the four skill models by our subject matter expert[27]. To predict a student's test score when a particular skill model is adopted, first, the fractional score the student can get on each individual item is computed, and then the "item-scores" are summed up to give a total score for the test. So how do we predict their state test item-score?

Given a student's learning parameters on all skills, and the exact test date of MCAS, we can calculate the probability of positive response from the student to an item tagged with any single skill. In the case that an item is tagged with more than one skill, I pick the skill that gives the lowest probability among all the skills that apply to the item[28] for that student (i.e. the hardest skill for the student). Thus, I obtain the probability of positive response to any particular item in the state test. In this approach, a student's probability of correct response for an item is used directly as the fractional score to be awarded on that item for the student. Then,

27 All the tagging was done after the MCAS items were released without any reference to the modeling process described in this paper.

28 We admit that there are other approaches dealing with multi-mapped items. For instance, using Bayesian Networks is a reasonable way to deal with this situation. Pardos et al. (2007) use this approach and got similar results that fine-grained models enable better predictive models.

item scores are summed up to produce the total points awarded on the test. For example, if the probability of an item marked with Geometry is 0.6, then 0.6 points will be added to the sum as the points awarded for this item. This sum is just what I use as the predicted state test score[29].

The prediction evaluation functions we build using the existing data are also intended to work well in future years, and so for reasons of interpretability, the prediction error function chosen is mean absolute deviation (MAD). A normalized metric named % Error is also calculated by dividing the MAD by the full score.

$$\%Error = MAD/(MaxRawScore)$$

where "*MaxRawScore*" is the maximum raw score possible with the MCAS questions used. The MCAS state test consists of 5 open response, 4 short answer and 30 multiple choice questions. The max score is 54 points if all 39 MCAS questions are considered, since some are scored wrong/right and some are scored with partial credit. In our case, only the multiple-choice and short answer questions are used with regard to the fact that currently open response questions

29 I think it might be useful to discuss our model from a more qualitative point of view. Is it the case that if you tag an item with more skills, does that mean our model would predict that the item is harder? The answer is no, in the sense that if you tagged a bunch of items with an easy skill (i.e., one easier than what the item was currently tagged with), that would not change our model's prediction at all. This makes qualitative sense, in that we believe the probability of getting a question correct is given by the probability of getting correct the most difficult skill associated with that question.

are not supported in our system. This makes a full score of 34 points with one point earned for a correct response on an item.

5.3.4 Research question 1: Does adding scaffolding responses help identifying weak skills?

Given the data, I start the work examining only students' responses to original questions. The first research question I want to answer is: *Would adding response data to scaffolding questions help us do a better job of identifying students' weak knowledge and thus more accurately predicting state test scores, compared to only using the original questions?* Because the scaffolding questions break the test question down into a series of simpler tasks that directly assess fewer knowledge components, which allows us to deal with skill "credit-and-blame", I hypothesize the ASSISTments System can do a more accurate job assessing students' skill mastery level. This hierarchal breakdown of knowledge provides a much finer-grained analysis than is currently available. It creates a good chance for us to detect exactly which skills are the real obstacles that prevent students from correctly answering the original questions. I also hypothesize that scaffolding questions would be especially useful when we utilize a finer-grained model since multi-mapping of original questions becomes more often in such a model and thereby "identifiablity" becomes a more serious issue. I think that getting an answer to this question would help us properly evaluate the second and more important research question described in Section 5.3.3.

5.3.4.1 Scaffolding credit and partial blame

Since the scaffolding questions show up only if the students answer the original question incorrectly, their responses to the scaffolding questions are only explicitly logged in that situation. However, if a student gets an original question correct, he is only credited for that one question in the raw data. To deal with the "selection effect," I introduce a compensation strategy of "scaffolding-credit": all associated scaffolding questions are also marked correct if the student gets an original question correct, assuming they have mastered every single knowledge component required by this problem.

An important thing we need to determine when using a multi-mapping model (in which one item is allowed to be tagged with more than one skill) is which skills to blame when a student answered an item tagged with multiple skills incorrectly. Intuitively, the tutor may want to blame all the skills involved; however, this would be unfair to those relatively easy skills when they are tagged to some compound, hard items. To avoid this problem, I apply the "partial blame" strategy: if a student got such an item wrong, the skills in that item will be sorted according to the overall performance of that student on the skills and only the hardest skill for the student (i.e. the skill on which that particular student showed the worst performance) will be blamed.

When evaluating a student's skill levels, both original questions and scaffold responses are used in an equal manner and they have the same weight in evaluation.

5.3.4.2 Results and discussion

Recall that the research question asks whether adding response data to scaffolding questions can help us do a better job of tracking students' knowledge and more accurately predicting state test scores. To answer the question, I first train mixed-effects logistic regression models using the data set that only includes original questions response; one regression model for each skill model. Then I replicate the training process. This time student responses to scaffolding questions are included and the "credit-and-blame" strategy described as above is applied. Again models are trained for all four skill models.

Interestingly but not surprisingly, it turns out that better-fitted models, as measured by %Error, on the state test can always be obtained by using scaffolding questions. In particular, when using the WPI-1 on DATA-2005, the mean decrease of % Error is 1.91% after scaffolding questions are introduced; for WPI-5, the decrease is 1.21%; and the decrease of % Error is 2.88% for the WPI-39 and 5.79% for the WPI-78 which is the biggest improvement. I then run paired t-tests between the % Error terms for the 447 students and find that the improvements are statistically significant in all the four cases as summarized in

Table 5.2. The same effect shows in DATA-2006. As presented in Table 5.2, the improvement on %Error is statistically reliable on all of the four models.

Table 5.2. The effect of using scaffolding questions on DATA-2005 and DATA-2006

Skill Model	MAD		% Error (MAD/#items)		Δ %Error	p-value of paired t-test
	Orig. Response	Orig.+ Scaffolding Response	Orig. Response	Orig.+ Scaffolding Response		
DATA-2005						
WPI-1	5.07	4.42	14.91%	13.00%	1.91%	0.008
WPI-5	**4.78**	4.37	**14.06%**	12.85%	1.21%	0.049
WPI-39	5.20	4.22	15.29%	12.41%	2.88%	<0.0001
WPI-78	6.08	**4.11**	17.75%	**12.09%**	5.79%	<0.0001
DATA-2006						
WPI-1	6.81	6.58	20.05%	19.37%	0.63%	0.001
WPI-5	6.76	6.51	19.88%	19.14%	0.74%	<0.0001
WPI-39	5.98	4.83	18.68%	15.10%	3.58%	<0.0001
WPI-78	5.58	4.99	16.91%	14.70%	2.21%	<0.0001

This drop-down of %Error (also MAD) makes sense for several reasons. One is that by using the response data to scaffolding questions I am using more of the information we collected. So it is not surprising that I can get a better model. A second reason is that the scaffolding questions help us do a better job of dealing with credit-and-blame problems. The effect of scaffolding questions is so strong that the order of the models even shift for DATA-2005. When only original questions are used, the WPI-5 is the best model (as bolded in Table 5.2), and the prediction error gets worse when the WPI-39 or WPI-78 models are used; while after scaffolding question responses are introduced, the most fine-grained model, WPI-78 becomes the best one. My interpretation of this is that when only original responses are used, individual skills do not get as much identifiability; it only makes sense to make fine-grained skill models, if you have questions that can be

tagged with just a single skill. Another reason why finer-grained models might not fit the data as well would be the fact that the finer-grained model has fewer data points per skill, so there is a tradeoff between the number of skills you would like and the precision in the estimates. For instance, for DATA-2005, the amount of data points for each skill in the WPI-39 is about 2.3 as many as data for those skills in the WPI-78. I also look at the "branching factor" (i.e. on average how many skills are tagged with each question) of different models for both data sets. In DATA-2005, branching factors of original questions are 1.13, 1.28 and 1.35 respectively for the WPI-5, WPI-39 and WPI-78 models. The factors decrease to be 1.006, 1.03 and 1.05 for scaffolding questions. In DATA-2006, also for the three models in the same sequence, branching factors of original questions are 1, 1.37 and 1.42, and the numbers are 1, 1.12 and 1.04 for scaffolding questions. So it is true in these data sets that the finer-grained models tend to be more frequently multi-mapped; and scaffolding questions are usually tagged with one skill, occasionally two. To get more "identifiability" per skill, in the next section I use the "full" response data (with scaffolding question responses added in) to try to answer the question of whether finer-grained models predict better.

Sharp readers may have noticed that the MAD of WPI-39 model for DATA-2006 is lower than that of WPI-78, yet %Error of the WPI-39 model is higher than %Error of the WPI-78 model. This is because the two multiple-choice items in 2006 MCAS test, item 13 and item 26, were tagged with the skills "N.6.8-

understanding-absolute-value" and "P.9.8-modeling-covariation" respectively, yet, none of the ASSISTment System items were tagged by the same two skills, which means we do not have training data to track student knowledge on the two skills. Therefore, I ignored the two items when predicting students' total score of 2006 MCAS test using the WPI-39 model. This reduces the total number of MCAS items of the WPI-39 to 32. The %Error of the WPI-39 model is calculated by MAD/32 while the %Error of the other models is calculated by MAD/34.

Does an error rate of 12.09% on the WPI-78 seem impressive or poor? What is a reasonable goal to shoot for? Zero percent error? For comparison reason, we created a baseline estimation of students' MCAS test scores by first computing students' overall percent correct on original questions, and multiplied the %correct with the full score. Under this "dumb" approach, the %Error was 17.26% for DATA-2005, and 21.47% for DATA-2006. In Feng, Heffernan & Koedinger (2006b) we reported on a simple simulation on how well one MCAS test was at predicting another MCAS test. We did not have access to data for a group of students who took two different versions of the MCAS test to measure this, so we estimated it by taking students' item level scores on MCAS, randomly splitting the 34 multiple-choice items in the test into two halves, and then using their scores on the first half to predict the second half. This process was repeated fives times, and on average the %Error was 11%, suggesting that a 12% error rate is looking somewhat impressive.

5.3.5 Research question 2: Does the finer-grained model predict better?

The second and more important research question I want to answer has to do with how the granularity of a skill model affects its effectiveness on tracking student knowledge. Specifically, the question is: *How does the finer-grained skill model (WPI-78) do on estimating external test scores compared to the other skill models?* We think that an answer to this question that says that a finer-grained model allows for better modeling/prediction would have important societal implications (e.g., regarding tracking student performance and reporting to teachers).

5.3.5.1 Does the fine-grained model fit better than the coarser-grained models?

To answer RQ2, I compare the four mixed-effects regression models (trained on the "full" data set with scaffolding questions used) fitted using the four different skill models. As shown in Table 5.3, the WPI-78 gets the best result, followed by the WPI-39, WPI-5, and followed by the WPI-1. % Error drops down when a finer-grained model is used, from WPI-1 to WPI-5 and then from WPI-39 to WPI-78.

In order to see if the % Error is statistically significantly different for the models, I compare each model with every other model. First, 95% confidence intervals are also calculated for MAD of all models. We notice although the p-

value from paired t-test are small, there is overlap among the confidence intervals. Generally speaking, the confidence intervals are very wide and the standard deviations of MAD are big. So the prediction accuracy varies across students. Then, paired t-tests are conducted between individual % Error term of the 447 students in DATA-2005 and also the 474 students in DATA-2006. I find out that in DATA-2005, the WPI-78 does as well as the WPI-39 ($p = 0.21$), and they both predict MCAS score reliably better than the WPI-5 and WPI-1. In DATA-2006 the WPI-78 model is statistically reliably better than the WPI-39, WPI-5 and WPI-1 ($p<0.001$ in all cases), and WPI-1 is statistically reliably worse on predicting MCAS scores than the other models ($p <0.0001$), suggesting that finer-grained skill models are more helpful in tracking students' knowledge over time.

Table 5.3. Evaluating the accuracy of MCAS prediction across skill models

Skill Model	MAD	95% Confidence Interval for MAD	% Error	
DATA-2005				
WPI-1	4.42	[4.12, 4.72]	13.00%	p=0.006
WPI-5	4.37	[4.07, 4.66]	12.85%	p<0.0001
WPI-39	4.22	[3.94, 4.50]	12.41%	p=0.21
WPI-78	4.11	[3.84, 4.39]	12.09%	p=0.10
IRT-2005	4.36	[4.04, 4.68]	12.82%	
DATA-2006				
WPI-1	6.58	[6.18, 6.99]	19.37%	p<0.0001
WPI-5	6.51	[6.11, 6.90]	19.14%	p<0.0001
WPI-39	4.83	[4.56, 5.11]	15.10%	p=0.0001
WPI-78	4.99	[4.71, 5.28]	14.70%	p=0.03
IRT-2006	4.67	[4.34, 4.99]	13.7%	

As a measure of internal fit, the average absolute residuals for each model fitted on the data are also calculated. For data of both years, the WPI-78 fits best. Since the WPI-78 model contains far more skills than other models, one might think the model wins simply because of the large number of parameters. As a sanity check, we generate a Random-WPI-78 model in which items are randomly mapped with skills from the WPI-78 model. It turns out that the random model does reliably worse than the WPI-78 model (and also the WPI-39), both in MCAS score prediction and in the internal fit[30].

I want to address that our results on student performance prediction are by no means the best. As a matter of fact, an Item Response Theory (van der Linden & Hamilton, 1997) model that has been widely used in traditional testing area by psychometricians as a control has also been trained. I fit the simplest model on the same online data as being used above (both scaffolding and original response data after pre-processing), the Rasch model that models student *i*'s dichotomous response (0 = wrong, 1 = correct) to problem *j* as a logistic function of the difference between student proficiency (θ_i) and problem difficulty (β_j), on our

[30] It is common to report the value of a model by using a metric that balances model fit and model complexity such as Bayesian Information Criterion (BIC). For instance, Cen et al. (2005) and Ferguson et al. (2006) both used BIC to compare different models. However, because the size of the datasets was different when we used the different models; the finer-grained models add additional rows for all questions that are tagged with more than one skill while BIC only make sense when the data is meant to be the exact same size. For the same reason, we did not conduct ANOVA on the results.

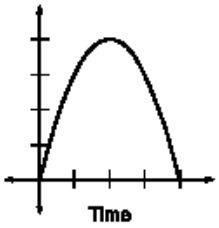
online data. The fitted model give us an estimate of math proficiency for every individual student which allows us to compute the predicted MCAS score assuming every item in MCAS has an average difficulty (β=0). In Table 5.3, *IRT-2005* refers to the IRT modeling condition for DATA-2005, and *IRT-2006* refers to the IRT modeling for DATA-2006. As we can see, the %Error of the Rasch model for DATA-2005 is 12.82%, marginally higher than that of the WPI-78, 12.09% ($p = .10$). Yet, the Rasch model does better in the next year where the %Error (13.70%) is reliably higher ($p = .03$) than that of the WPI-78 (14.70%). Other than the IRT model, we have also contrasted our result on DATA-2005 with the result produced by Bayesian network approach that dealt with skills associated with one item conjunctively using "AND" gate (Pardos, Feng, Heffernan, & Heffernan, 2007). The "AND" gate signifies that all the skills must be known in order for the questions to be answered correctly. Pardos et al. (2008) confirmed the "conjunctive" hypothesis. During the comparison process, we found out that our approach did better than the Bayesian networks approach when the WPI-1, and WPI-5 models were used, and the two approaches are comparable when the WPI-39 and WPI-78 were used. Specifically, for the WPI-39 model, %Error of the Bayes approach is 12.05%, lower than what we got (12.41%); yet for the WPI-106 model, %Error the Bayes approach is 13.75%, higher than our result of 12.09%.

Comparing the results I get using DATA-2005 and those using DATA-2006, I notice two things changed. First, the order of prediction accuracy differs when only original questions are used. The finer-grained models still track student knowledge better than coarser-grained models when DATA-2006 was used; yet it is not the case when DATA-2005 was used. Second, the prediction error is much higher in the year 2005-2006 than in the previous year. Third, the effectiveness of the IRT model reduces in the year 2006. One possible reason for the higher prediction error is that we have fewer training data points for each student in the year 2005-2006. On average, in the year 2005-2006, each student was involved in 5.5 sessions and finished 51 problems while in the year 2004-2005, each student got on average 7.3 data points and did 87 problems. Additionally, the problem set administered to students in the two years are not the same, and the amount of questions included differ as well. The data set from DATA-2005 includes 329 original questions, and 887 associated scaffolding questions. On the contrary, DATA-2006 only includes 222 original questions and 354 scaffolding questions, with each original questions associated with less scaffolding questions than in DATA-2005.

5.3.5.2 How well does the fine grained model fit the data?

When using logistic regression, the statistical packages allow the user to analyze which of the parameters seem to have good fitting values. We now turn to

do a little more close analysis on the WPI-78 to see how good our model is. In our model, each skill gets one coefficient indicating the skill's "intercept" and one for the skill's "slope." The first of these, the intercept, allows us to model that some skills start the year with students knowing them better, while the slope allows for the fact that some skills are learned more quickly than others. The model shows that for students who used the system in the school year 2004-2005, the easiest skills are "Subtraction," "Division" and "Simple-Calculation," while the skill that had the hardest incoming difficulty was "Qualitative-Graph-Interpretation" (as shown in Figure 5.3). I also look at the fits on the slopes for each skill. The skill that shows the steepest rate of learning during the course of the year is "Sum-of-Interior-Angles-Triangle" (e.g., "what is the sum of the angles inside of a triangle?"). It seems quite plausible that students learned a good amount related to this skill as we noticed in a classroom a poster that said the "The sum of the interior angles in a triangle is 180" clearly indicating that this was a skill that teachers were focused on teaching. The skill that showed the least learning is called "Equation-Concept" (as shown in Figure 5.4). Out of the 78 skills, 7 coefficients predicted "un-learning" (i.e., the slopes are negative), which presumably raises a sign of overfitting, or that the tagging of the skills in the skill model is not quite right.



Which of the following could be shown by the graph?

- the height of a candle as it burns over time
- the distance covered by a car traveling at a constant speed over time
- the height of water in a tank being drained at a constant rate over time
- the height of a ball thrown straight upward over time

Figure 5.3. A question tagged with the skill "Qualitative-Graph-Interpretation."

How many cylinders must be placed on the empty side of the second scale to make that scale balance?

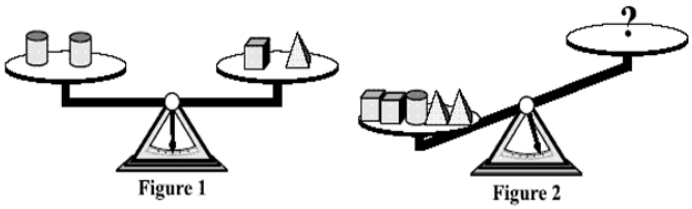


Figure 5.4. A question tagged with the skill "Equation-Concept"

Speaking of the accuracy of fit, I also notice that the model obtained a high accuracy on predicting student response on items tagged with the simple skills (e.g. Division, Subtraction), yet not so good at tracking student knowledge on skills "Of-Means-Multiply", "Interpreting-Linear-Equations" or "Inequality-Solving". I speculate that skills that have less data for them would be more likely to be poorly fit. I calculate the correlation between the total amount of data points

for each skill and the average residual of the skill, and find out that the correlations are weak. Other reasons a skill might have a poorly fit slope would be that we tagged items with the same skill names that share some superficial similarity, but do not have the same learning rates. This analysis suggests some future work in refining the WPI-78 model; for instance, one possible refinement is to merge "equation-concept" with "equation-solving" (i.e., delete the "equation-concept" skill from the model and map all items tagged with "equation-concept" to "equation-solving").

All in all, I make no claim that the fine grained model that we created represents the best fitting model possible. Nevertheless, I stand by the claim that this model, taken in total, is good enough that it can produce good fit to the data, and make good predictions of the MCAS scores, indicating the model is useful, even given the flaws that might exist in it.

5.4 Related Work and Contribution

There has also been a large interest in building cognitive diagnostic models. What we refer to as a "skill model" is referred to as "Q-Matrix" by some educational data mining researchers (Barnes, 2005, 2006) and psychometrician (Tatsuoka, 1990); Croteau, Heffernan & Koedinger (2004) called it "transfer model"; while Cen, Koedinger & Junker (2005), and Gierl, Wang & Zhou (2008)

used the term "cognitive model." In all cases, a skill model is a matrix that relates questions to the skills needed to solve the problem. Such a model provides an interpretative framework to guide test development and psychometric analyses so test performance can be linked to specific cognitive inferences about the examinees.

Researchers in machine learning area have been using automatic/semi-automatic techniques to search for skill models. Tatsuoka and colleagues developed the rule space method (Tatsuoka, 1990, 1993) in which hypothesized expert rules and actual student errors in fraction addition can be mapped and compared. The expert point that is closest to the student response is assumed to be the rule that the student is using. Barnes has done considerable work with trying to induce transfer models, in this work called Q-matrices (Birenbaum, Kelly, & Tatsuoka, 1993), from data (Barnes, 2005, 2006). Koedinger and colleagues (Koedinger & Junker, 1999; Cen, Koedinger & Junker, 2005, 2006) proposed a semi-automatic approach called Learning Factor Analysis (LFA) as a generic solution to evaluate, compare, and refine many potential cognitive models of learning. Pavlik, Cen, & Koedinger (2009) proposed a method called learning factors transfer analysis to automatically generate domain models. Though addressing the same problem, my work is different from above in that we hand-coded the skill models and built the connection between skills and questions. This is similar to what Ferguson et al. (2006) did in their work as they also

associated problems with skills by hand, but they employed a different methodology using Bayesian Networks.

Though different approaches have been adopted to develop skill models and model students' responses, as far as I know, very little effort has been put in comparing different grain-sized skill models in the intelligent tutoring system area. The only work we are aware of that shows that by building fine-grained skill models researchers could build better fitting models was by Yun, Willett and Murnane (2004) who developed an alternative curriculum framework. Their results of confirmatory factor analysis showed that the alternative framework fits data better suggesting the state's learning standards is subject to improvement. However, they did not try to answer the question regarding to the right grain size of skill models. Collins, Greer, & Huang (1996) investigated the results of skill hierarchies. Yet, different from us, they were using simulated users, but have not applied their approach on real student data. Carmona1 et al. (2005) also introduced a hierarchal structure into their Bayesian models. However, they focused on prerequisite relations among skills but not the various granularities of skill models, as I do.

Designed to measure specific knowledge states and cognitive processing skills in a given domain, cognitive diagnostic modeling based on student response data has a long history (e.g. Nichols, Chipman & Brennan, 1995; Leighton, & Gierl,

2007). Corbett and Anderson (1995) employed a diagnostic approach called *knowledge tracing* that models students as an overlay of the ideal production rules. Corbett and his colleagues (Corbett, Anderson, & O'Brien, 1995) employed a very detailed model of skills, but their system did not have questions tagged with more than one production rule (Anderson, 1993). However, in this dissertation, multi-mapping[31] is allowed. I propose very simple strategies to take advantage of the structures of ASSISTments to deal with the credit and blame issue, which turns out to be as effective as the complex conjunctive modeling approach.

Anderson (1982) and Newell & Rosenbloom (1981) found that the performance of cognitive skills improves approximately as a power function of practice. Draney, Pirolli, & Wilson (1995) followed the power function form of learning, but decomposed the parameter related to the first trial performance in the power function into effects with respect to production rules difficulties, pedagogical manipulations, and individual difference. Our collaborators (Ayers & Junker, 2006) were engaged in trying to allow multi-mapping using a version of the fine grained model but reported their Linear Logistic Test Model (LLTM) does not fit well. Different from my approach, the model Ayers & Junker applied does not track student performance over time. Also, my method is different

[31] A "multi-mapping" skill model, in contrast to a "single-mapping" model, allows one item to be tagged with more than one skill.

Anderson and Draney's in that the mixed-effects model assumes linear learning and *time* instead of trial is used as a covariate. Since students only use ASSISTments sparsely across a year, using *time* accounts for the learning happened during normal classroom instruction. For instance, in my model, a response given in October will be treated differently from a response given in December, even if both represent a student's second trial on a math skill "Pythagorean theorem". Yet, the power function of learning would consider no difference.

Mixed-effect models are used in this dissertation to track student knowledge development. In the literature, many other approaches have been developed to address the same problem as well. Gierl, Wang & Zhou (2008) used an approach called "attribute hierarchy method", a cognitively-based psychometric procedure developed by Leighton, Gierl, and Hunka (2004) at the University of Alberta, make diagnostic inference about cognitive skills in SAT. Mislevy and his colleagues are using a methodology they call "Evidence Centered Design (ECD)" in their approach to cognitively diagnostic assessment at Educational Testing Service (Mislevy, Steinberg, & Almond, 2003). The mapping between the knowledge components is represented in a Bayesian Network. Almond et al. (2007) also examines the application of Bayesian networks to Item Response Theory-based cognitive diagnostic modeling. In these works, the Bayes net does the inferencing to estimate each student's knowledge of each component. This

work is widely respected in psychometrics while less known in the intelligent tutoring systems community. A major difference from my work is that these approaches have usually been used on cases where students are not learning during (or between) test sessions. In the moving target circumstance, Dynamic Bayesian Networks (Dean and Kanazawa, 1989) would be a natural extension to apply to the ECD method when students are learning. In the tutoring situation, VanLehn and colleagues (e.g. Murray, VanLehn, & Mostow, 2001) have applied a similar approach.

Others (e.g. Bock, Gibbons, and Muraki, 1988), in the psychometrics field, have developed multi-dimensional Item Response Theory (IRT) models but these models do not allow multi-mapping. Different from most other efforts (including studies from this dissertation) focusing on correct knowledge, Baffes & Mooney (1996) implemented an approach in an authoring system called ASSERT which uses theory refinement to introduce errors into an initially correct knowledge based so that it models incorrect student behaviors, and generate feedback for remediation.

5.5 Conclusion

Mislevy (2006) described six steps in model-based reasoning in science. These steps, including model formation, elaboration, use, evaluation, revisions and

model-based inquiry, provide a framework for considering my progress in developing and refining skill models. In this chapter, I establish evidence that we can over time, track students' skills in an intelligent tutoring system like the ASSISTments System. I find solid evidence that using students' responses to scaffolding questions were helpful in tracking students' knowledge and fine-grained models can better predict MCAS scores than coarser ones. As discussed in the chapter, teachers want reports by skills at fine-grained level, and this is the first evidence we have saying that our skill mappings are good enough to better predict a state test than some less fine grained models. Furthermore, I illustrate how individual skills are modeled and being used to give feedback to teachers in the ASSISTments system. I argue that the results show further that not only can reliable assessment and instructional assistance be effectively blended in a tutoring system, but also, more importantly, such a system can provide teachers with useful fine-grained student-level knowledge they can reflect on and adjust their pedagogy.

Chapter 6

6 Towards Improving System Effectiveness

Chapter organization: This chapter explores a mixture of efforts on improving the overall effectiveness of a tutoring system. After an introduction is given in Section 1, Section 2 focuses on detecting instructional effectiveness of the tutoring system by answering two research questions. Given the results from Section 2, Section 3 describes how these findings can be used to help manually or automatically improve existing skill model. Section 4 illustrates a different way to refine existing models, aiming at increasing their power of modeling students at different knowledge levels. Section 5 highlights the contribution of the chapter and reviews related works. Section 6 concludes the chapter. The essence of this chapter has been included in the following papers: Feng, Heffernan, Beck, & Koedinger, 2008; Feng, Heffernan, & Beck, 2009 (Section 2); Feng, & Beck, 2009 (Section 4). I want to point out that the work described in Section 3 of this

chapter is the only work in this dissertation that is in preparation and has yet been published.

6.1 Introduction

In Chapter 4 and Chapter 5, I have argued that the ASSISTments system can accurately predict student end-of-year performance, and a fine grained skill model is developed to track student's mastery of individual knowledge components and to inform teachers more precisely. Based upon above works, in this chapter, I will focus on improving the system, including overall tutoring efficacy and knowledge modeling accuracy. Below I will illustrate my efforts on trying to improve the instructional effectiveness through several steps. First, applying educational data mining techniques to find out whether a system effectively teaches; secondly, detecting if there is variance of the instructional efficacy among different skills; thirdly, more closely examining if the variance exists among individual pieces of instructional contents.

Once I can reliably tell difference among skills and items, this information can be utilized to improve existing skill models which in turn improve the overall predictive power of the system and thus could potentially increase learning. Two approaches are proposed for this purpose. First, support subject experts' efforts on manually update existing hand-crafted skill models by providing our findings

from data analysis. Secondly, I also try to suggest difficulty factors for learning factor analysis based on the research findings on item effectiveness variance so that LFA can automatically searches for superior skill models.

As a stream of the efforts of refining existing skill models, several works have been done to use automatic and computationally intensive methods to search for best-fit models for groups of different knowledge levels. Yet, I have taken an alternative view to start from our hand-constructed skill models and use them as a lens to examine student learning. This work follows Chapter 5 in that it examines the models based on their granularities. Yet it also reflects my efforts on refining existing models. I thereby describe this work in this chapter as opposed to organizing it into Chapter 5.

6.2 Using item level analysis to analyze instructional effectiveness

The field of Intelligent Tutoring Systems is often concerned with how to model student learning over time. More often than not, these models are concerned with how student performance changes while students are using the tutor. In Chapter 4, I have described my work on tracking student learning longitudinally over time and also pointed out that it confounds student's learning in the classroom and also within ASSISTments. While in this section, I will look

to see if learning from the computer system is happening over time, trying to separate out learning from the classroom. For these purpose, I conduct a focused item-level analysis of a subset of items to track how student performance on these items changes during the same ASSISTment session. I will explore the possible reasons of why on some sets of problems students learned or failed to learn. And on top of that, I want to see if we can tell which item in a group is the most effective at causing learning.

6.2.1 Description of the data

The hypothesis is that students learn from groups of items that share the same background knowledge requirement. Our subject manner expert picked 181 items out of the 300 8th grade (approximately 13 to 14 years old) math items in ASSISTment. Items that have the same deep features or knowledge requirements, like approximating square roots, but have different surface features, like cover stories, were organized into a **Group of Learning OPportunity (GLOP).** Besides, the expert excluded groups of items where learning would be too obvious or too trivial to be impressive. Also, GLOPS had to be of at least size two, so singleton items were not selected either. The selected 181 items fall into 39 GLOPs with the number of items in each GLOP varies from 2 to 11. The items are a fair sampling of the curriculum and cover knowledge from all of the five major content strands identified by the Massachusetts Mathematics Curriculum

Framework: Number Sense and Operations; Patterns; Relations and Algebra; Geometry, Measurement, and Data Analysis; and Statistics and Probability. It sampled relatively heavily on the strand Patterns, Relations & Algebra. Items in the same group were collected into the same section of ASSISTments, and seen in random order by students. Each student potentially saw 39 different GLOPs that involve different 8th grade math skills (e.g. fraction-multiplication, inducing-functions, symbolization articulation) in random order. Figure 6.1 shows four items in one GLOP that are about the concept "Area." All these problems asked students to compute the area of the shaded part in the figures. It is worth pointing out that all the GLOPs were constructed by focusing on the content of the items before the analysis is done in this dissertation.

I collected data for this analysis from Oct. 31, 2006 to Oct. 11th, 2007. Over 2000 8th grade students participated in the study. Cases where the student only finished one item in a GLOP are excluded. We end up with a data set of 54,600 rows, with each row representing a student's attempt at an item. 2,502 students enter into the final data set, mostly from Worcester, Massachusetts area. Each student on average worked on 22 items.

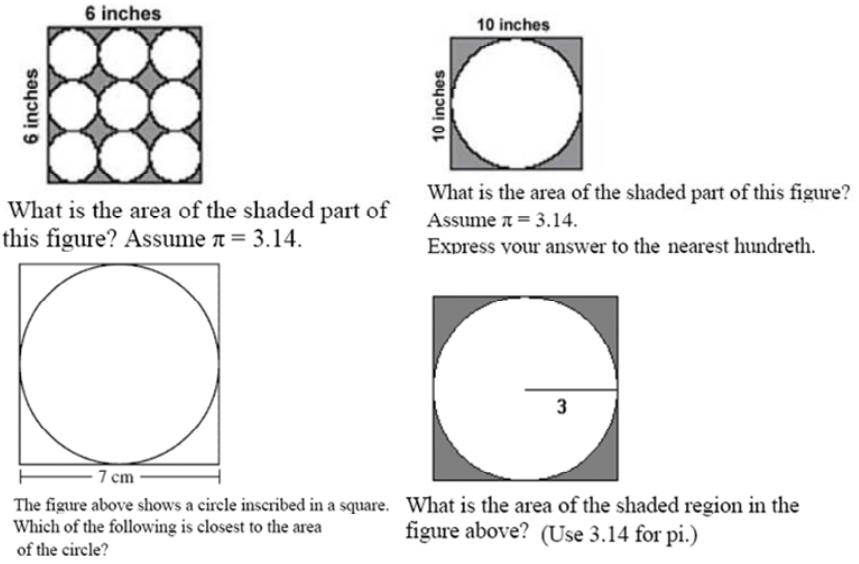


Figure 6.1. A sample GLOP that addresses the skill "Area"

Although one may argue for other indicators, e.g. students' help requests and response times, for this study, I simply choose to use the correctness of student's first attempt to an item as an outcome measure of their performance. Table 6.1 shows a sequence of time-ordered trials of student "Tom" and "Mary" on items in two GLOPs, together with the correctness of each response. The column "correct?" indicates whether the student answered the question correctly or not. The value will be 1 where he succeeded; otherwise, it is set to be zero. "Tom" finished all 5 items in the GLOP 1. He managed to solve the first, the forth and the fifth item but failed on the second and the third one. "Mary" worked on four problems (i.e. had 3 opportunities to learn) from GLOP 4 on April 2nd, 2007 starting from 9:39 AM. She failed the first two but successfully solved the last two items.

Table 6.1. Raw response data of two students on two GLOPs

Student	GLOP	Item ID	Timestamp	Previous trials (t)	Correct?
Tom	1	1045	11/7/2007 12:30:31AM	0	1
Tom	1	1649	11/7/2007 12:31:15 AM	1	0
Tom	1	1263	11/7/2007 12:43:40 AM	2	0
Tom	1	1022	11/7/2007 12:46:09 AM	3	1
Tom	1	1660	11/7/2007 12:48:20 AM	4	1
Mary	4	2236	4/2/2007 9:39:20AM	0	0
Mary	4	9086	4/2/2007 9:42:19AM	1	0
Mary	4	2239	4/2/2007 9:53:11AM	2	1
Mary	4	2274	4/2/2007 9:58:07AM	3	1

6.2.2 Research question 1: Can we predict from which groups of questions students will learn?

6.2.2.1 Do students learn from ASSISTments?

I first attempt to determine whether the system effectively teaches (Feng, Heffernan, Beck, & Koedinger, 2008). Learning is assessed by comparing student performance the first time they were given one item from a GLOP with their performance when they were given more items (also more opportunities) from the same GLOP in the same day. If students tend to perform better on later opportunities of items in a GLOP, it indicates that they may have learned from the instructional assistance provided on items by the ASSISTments system that they worked on earlier by answering the scaffolding questions or by reading hint messages. There is controversy over whether same-day learning opportunities should be used as evidence of learning. For example, Beck (2006) thought repeated trials were not indicative of learning. He chose not to use later encounters on the same day in the Reading Tutor since performance on those encounters is not a reflection of student knowledge but just retrieved from short

term memory. However, our domain (mostly 8th grade multi-step math problems) is more complex than reading and the items in a GLOP usually have different surface features. Solving these problems is not simple retrieval of an answer from a previous question. And even if it isn't more complex, our later day trials are horribly confounded by classroom instruction due to low density of usage (every other week). Therefore, I choose to analyze the response data on the same day to eliminate the confound of learning happening because of classroom instruction **between** two ASSISTment sessions.

To answer the question, I run a logistic regression to study the relationship between student performance (i.e. their responses to items) and the number of opportunities the student has on a GLOP. In this method, the dependent variable is student response to a question and we account for the difference of student math proficiency by including *student* as one of the predictor variables. Similarly, *question* is included as another predictor with regard to the fact that questions in one GLOP may vary in difficulties. *Opportunity* is introduced as a covariate. The regression formula is

Equation 6.1. Logistic regression model

$$\ln(\frac{p_{ij}}{1-p_{ij}}) = \alpha_i * Student_i + \beta_j * Question_j + \gamma * Opportunity\#$$

Where

p_{ij} is the probability that the student *i* will answer question *j* correctly

Opportunity# indicates how many opportunities the student *i* has on a particular GLOP.

α_i, β_j and γ are the coefficients for the corresponding predictors $Student_i$, $Question_j$ and *Opportunity#.*

The model is very similar to models used by Cen et al. (2005) for Learning Factor Analysis except that skills are not included as factors since we are investigating generalized learning over all GLOPS. A multinomial logistic regression is run in SPSS and the regression coefficient estimated by the model, corresponding to the number of learning opportunities (γ), is .03 ($p < .001$). This result suggests that in general, students perform reliably better as they have more chances of practicing on the same GLOP. The coefficient tells that students will improve by 0.03, on a *logit* scale, for each practice opportunity. This learning corresponds to approximately a 0.8% improvement in performance for each problem practiced, a rather small effect. In Massachusetts, MCAS test scores are categorized into four performance levels (namely warning, need improvement, proficient, and advanced). According to the results of 2006 MCAS test, students need to earn 13 more points (24% of the full score) to jump from need improvement to proficient which is required by the federal movement based on NCLB standards to graduate from high school. Theoretically, if students can gain 0.7% for each learning opportunity, they will fulfill the 24% improvement by

solving 31 problems in ASSISTments. It should be noted that there may be a selection effect in this experiment in that better students are more likely to do more problems in a day and thereby more likely to contribute to this analysis. Also there is a limitation with the model that all GLOPs are assumed to produce the same amount of learning as indicated by the coefficient of the learning opportunity covariate, which may not be true.

Once I have found that students are learning from working in ASSISTments and the learning results are generalized across the 40 GLOPs, I then step further to explore if all of the GLOPs are equally effective at promoting learning. The answer is "no", which is not surprising anyway since the items in different GLOPs vary on several aspects (e.g. focusing on various skills; built by authors with differing teaching experience using various teaching strategies, etc.). In summary, out of the 40 GLOPs, the amount of learning per opportunity is statistically reliably higher than zero on 11 of them. 2 GLOPs caused marginally reliable learning and 16 caused unreliable learning. And there is non-reliable "un-learning" for the remaining 11 GLOPs, suggesting that not much learning occurred when students worked on these GLOPs.

6.2.2.2 Why students learned or failed to learn?

Now that I have shown that learning varies among GLOPs, I will explore the reasons for this variation. I am not only interested to know which category each

GLOP falls in and but also curious why. Particularly, I want to investigate why students did not show learning on certain GLOPs. Our four hypotheses are:

H1: Learning transfer from harder items to easier items, or students tend to learn more by doing harder items than by doing easier items. Presumably, if a student learned to solve a hard item, he then should be able to do better on an easier item that requires similar skills. However, the converse is not necessarily true.

H2: Knowledge transfer occurs within GLOPs of items that use similar skills. We can never know exactly how a student internally represents a problem and what the exact skills a student applied to solve a problem. But if a GLOP is well-focused in what it covers, presumably students should show more learning within it.

H3: The "learnability" of the skills required by GLOPs varies. Our statistics show that each ASSISTment provides about 2 minutes of instruction. It can be hard to teach some skills effectively, for instance, *symbolization articulation*, in such a short period. Such skills require deep understanding and more practice to be able to apply and transfer, whereas some other skills such as *area* are more teachable since students only need to be reminded to apply the area formula.

H4: The efficacy of instructions has an impact on learning results. We can easily imagine that some GLOPs have better teaching efficacy than others. The quality of the scaffolding questions and hint messages can differ from one item to

another as authors used a tutoring strategy that are more, or less, effective than others.

In this dissertation, I test the first two hypotheses and leave the last two as future work. The plan is to invite more content experts to help us identify the learnability of the related skills and to evaluate the quality of the ASSISTments by looking closely at the scaffolding questions and hint messages.

H1: Do students learn more from harder items or easier items?

Noticing learning varies among GLOPs, the first thing we do is to explore the relationship between the amount of learning and the easiness of a GLOP (measured by the average difficulty of items in the GLOP). I calculate the rank-order correlation and got a coefficient of .333 ($p = .036$, $N=40$), which indicates that students learned more on harder GLOPs than on easier ones. We asked ourselves: why is this? A quick answer is that there is more room to grow for harder items. Or, maybe students just learn more from harder items than easier items.

Beck (2006) introduced an approach called *learning decomposition* to analyze what type of practice was most effective for helping students learn a skill. The approach uses regression to determine how to weight different types of practice opportunities relative to each other. Here, learning decomposition is applied to our data set to investigate how students acquire math skills: will their

practice on harder items produce more learning? To test the first hypothesis, two more columns are added to the data set. One column, entitled "easier_before_current", represents how many items the student has seen in the same GLOP are easier than the current item. The other column, entitled "harder_before_current", indicates how many items were seen that are harder than the current one. I measure the easiness of the items using the item parameter given by a one-parameter Item Response Theory model (i.e. Rasch model[32]). The Rasch model was trained over data collected in the system from Sept., 2004 to Jan., 2008, including responses to 2,700 items from more than 14,000 students. The two columns are introduced as covariates in the regression model.

Equation 6.2. Learning decomposition model for H1

$$\ln(\frac{p_{ij}}{1-p_{ij}}) = \alpha_i * Student_i + \beta_j * Question_j + \gamma_e * easier\_before\_current + \gamma_h * harder\_before\_current$$

Where γ_e and γ_h represents the coefficients for the two new covariates respectively.

[32] In the Rasch model, the probability of a specified response is modeled as a logistic function of the difference between the person and item parameter. In educational tests, item parameters pertain to the difficulty of items while person parameters pertain to the ability or attainment level of people who are assessed.

The model is fitted in SPSS 14.0 as well. In the results, I notice that the coefficients of the two covariates of *easier_before_current* and *harder_before_current* are very close to each other (γ_e = .032, and γ_h = .033). The coefficient for *easier_before_current* is fractionally but not reliably lower (p=.966), which suggests that students learn as much from easier items as from harder items, and thus the first hypothesis is rejected.

H2: Does more learning occur in GLOPs that are more focused?

H2 is different than H1 in that it believes that transfer occurs within GLOPs that have similar difficulty questions (and hence address similar skills based on our assumption). To test H2, we want to investigate the relationship between the amount of learning that happened in each GLOP and the cohesiveness of the GLOP in term of item difficulty and the skills that are needed to answer the items.

Two approaches are used to quantify the cohesiveness of the GLOPs. The first metric is an automated measure that comes from a computer modeling process based on the assumption that if two skills, A and B, are better modeled by a single skill, then practice on either A or B symmetrically transfers to the other. During the modeling process, for each GLOP, we compared the BIC of two models. The first model treats each question as having a separate difficulty. The second model treats all questions as having the same difficulty, and thus had *number_of_questions_in_GLOP-1* fewer parameters. Presumably, if the

cohesiveness of a GLOP is high, we should expect the second model to fit better on our data as measured by Bayesian Information Criteria (BIC) (or any model fitting criteria that penalizes for model complexity). We follow the same procedure and calculated the difference of BIC values between two models for each of the GLOPs.

The second metric is based on our subject-matter expert's ranking of the cohesiveness of the GLOPs. As requested by us, our subject-matter expert set the rating from 1 to 5. A fit of 1 or 2 means that the items are very different. A 3 means there are some flaws in the selection, a 4 means there are just a few inconsistencies and a 5 means they fit very well. According to the ranking of the expert, 18 GLOPs got a fit of 5, 10 were given a fit of 4, 7 GLOPs got 3 and the remaining 5 GLOPs scored 2.

After obtaining the two metrics, we continue to analyze the relationship between the cohesiveness of the GLOPs and the amount of learning that happened in each of them. First, I calculate the rank-order correlation between the automated metric and the amount of learning but fail to find a significant relation ($r = .13$, $p = .94$). I then discretize coherence into 3 coherence bins: high, medium and low and performed a one-way ANOVA to explore whether there are any differences in the amount of learning, but find no main effect ($F = .676$, $p = .515$). After that, I conduct the same analysis using the expert ranking of the

cohesiveness. The rank-order correlation between fit and amount of learning is equal to .322 ($p = .045$). Yet the ANOVA shows no main effect of fit ($F = 1.573$, $p = .213$). Further more, instead of using five groups, I merge all GLOPs with fit less than 5 into one category named "non-perfect-fit" as a contrast to the ones with "perfect-fit" and run an independent sample t-test to compare the mean between the two categories. The result suggests that there is statistically reliably more learning happening in GLOPs of perfect fit ($t = 2.311$, $p = .030$).

To complete the third side of the triangle of learning/automated coherence metric/expert ranking, I also compute the correlation between our two metrics of fit/cohesiveness. They do not correlate with each other ($r = -.198$, $p = .22$), which means that an automated measure and an expert's judgment differ. In conclusion, H2 is supported by the expert's judgment but not by the result of data mining.

To summarize, I find evidence that suggests there is learning within ASSISTments. More interestingly, I find that the learning differed across the groups of items. A variety of hypotheses is tested to explain this phenomenon. The automated approaches are unable to account for the variation. However, human expert judgments are predictive as to which groups of skills were learnable.

6.2.3 Research question 2: Can we tell which item is the most effective at causing learning?

One popular method of determining whether one type of instruction is more effective than the other, or whether one tutor is more beneficial than another on helping student learn a skill, is to run a randomized controlled study. A major problem with the controlled study approach is that it can be expensive. A study could involve many users (in each condition), be of considerable duration, and require the administration of pre/post tests. To address this problem, Beck (2006) introduced an approach called *learning decomposition*, an easy recipe to enable researchers to answer research question such as what type of practice is most effective for helping student to learn a skill. In this section, I apply learning decomposition to compare the instructional effectiveness of single tutoring components, leveraging data collected in the ASSISTments system (Feng, Heffernan, & Beck, 2009).

6.2.3.1 Methodology

Beck (2006) introduced the idea of learning decomposition. It extends the classic exponential learning curve by taking into account the heterogeneity of different learning opportunities for a single skill. The standard form of exponential learning curve can be seen in Equation 6.3. In this model, parameter *A* represents students' performance on the first trial; *e* is the numerical constant

(2.718); parameter *b* represents the learning rate of a skill, and *t* is the number of practice the learner has at the skill.

Equation 6.3. Standard exponential learning curve model

$$performance = A * e^{-b*t}$$

Equation 6.4. Learning decomposition model

$$performance = A * e^{-b*(B*t_1+t_2)}$$

The model as shown in Equation 6.3 does not differentiate different types of practice, but just counts up the total number of previous opportunities. In order to investigate the difference two types of practice (I and II), the learning opportunities are "decomposed" into two parts in the model in Equation 6.4, in which two new variables t_1 and t_2 are introduced in replace of *t*, and $t = t_1 + t_2$. t_1 represents the number of previous practice opportunities at one type I; and t_2 represents the number of previous opportunities of type II. The new parameter *B* characterizes the relative impact of type I trials compared to type II trials. The estimated value of B indicates how many trials that one practice of type I is worth relative to that of type II. For example, a B value of 2 would mean that practice of type I is twice as valuable as one practice of type II, while a B value of .5 indicates a practice of type I is half as effective as a practice of type II. The basic idea of learning decomposition is to find an estimate of weight B that renders the best fitting learning curve.

Equation 6.4 factors the learning opportunities into two types, but the decomposition technique can generalize to *n* types of trials by replacing t with $B_1*t_1 + B_2*t_2 + \ldots + t_n$. Thus, parameter B_i represents the impact of a type *i* trial relative to the "baseline" type *n*.

Various metrics can be used as an outcome measurement of student performance. For instance, Beck (2006) chose to model student's reading time since it is a continuous variable. When it comes to a nominal variable, e.g. dichotomous (0/1) response data, a logistic model should be used. Now learned performance, (i.e. *performance* in Equation 6.4), is reflected by odds ratio of success to failure. Equation 6.5 represents a logistic regression model for learning decomposition.

Equation 6.5. Logistic models for learning decomposition

$$performance = \frac{P(correct\_answer)}{P(wrong\_answer)} = A * e^{-b*(B*t_1+t_2)} = e^{a+b*(B*t_1+t_2)}$$

Equation 6.5 can be transformed to an equivalent form as below:

$$\ln(\frac{P(correct\_answer)}{P(wrong\_answer)}) = a + b*(B*t_1 + t_2)$$

Different approaches have been established to track students' progress in learning. Among these, one technique by Koedinger and colleagues is called Learning Factors Analysis (LFA) (Cen, Koedinger & Junker, 2005, 2006). LFA was proposed as a generic solution to evaluate, compare, and refine many

potential cognitive models of learning. Since student performance is often represented by a dichotomous variable, logistic regression models have been used as the statistical model for evaluation. Although both LFA and learning decomposition are concerned with better understanding student learning, and both use logistic models, they have different assumption. LFA assumes all trials cause the same amount of learning but the skills associated with each trial may vary, while learning decomposition assumes the domain representation is constant but different types of practice cause different amounts of learning.

The reminder of the section explores applying learning decomposition approach to answer questions about how students' acquisition of math skills is impacted by different instructional items, and various tutoring strategies.

6.2.3.2 Modeling

Table 6.1 has shown a sequence of time-ordered trials of student A on items in GLOP 11, together with the correctness of each response. The student finished all 5 items in the GLOP. He managed to solve the first, the forth and the fifth item but failed on the second and the third one. So, for teaching the math skill involved in GLOP 1, which item is likely to be more (or less) effective for student proficiency development? In order to answer this question, we adopt the idea of learning decomposition. Each item in a GLOP is considered as a different type of practice; then students' practice opportunities are factored into practice at each

individual item. Since each student has at most one chance at an item, the number of previous opportunities at each item is either 0, indicating the student has not worked on that item, or 1, indicating the student has finished that item before. Table 6.2 shows the corresponding data after the trials are decomposed into component parts. Rather than counting the number of previous encounters, we instead count the number of prior encounters of each item in the GLOP.

Table 6.2. Decomposed response data of student "Tom"

Student ID	Item ID	Trials (t)	Correct?	Prior encounters				
				Item 1022	Item 1045	Item 1263	Item 1649	Item 1660
Tom	1045	0	1	0	0	0	0	0
Tom	1649	1	0	0	1	0	0	0
Tom	1263	2	0	0	1	0	1	0
Tom	1022	3	1	0	1	1	1	0
Tom	1660	4	1	1	1	1	1	0

Given the data in Table 6.2, in order to determine the influence of each item on student learning we use a logistic regression model. I use a logistic model since our data are dichotomous. By fitting a logistic regression model, I seek to model the odds of giving a correct answer as

Equation 6.6. Logistic regression model for examining effect of practice on different items

$$\frac{P(correct\_answer)}{P(wrong\_answer)} = \exp(A + \sum_{i \in D} B_i * t_i)$$

Here A is the intercept of the regression model. The remainder part, $\sum_{i \in D} B_i * t_i$, represents a learning decomposition model that simultaneously estimate the impact of all items in a GLOP. *D* is a space of all items in a GLOP. The space is

different for different GLOPs. Coefficient B_i represents the amount of learning caused by item *i* (or learning rate of item *i*). Generally, a positive estimate of B_i suggests students tend to perform better on later opportunities after they encountered item *i*; or students have learned from the instructional assistance provided on item *i* by the ASSISTment system that they worked on earlier by answering the scaffolding questions or by reading hint messages. t_i represents the number of prior encounters of item *i*. Note that, the t_i's account for all possible trials, and are thus equal to *t*. When the model parameters are estimated from data, the *B* parameters indicate the relative impact of different items on student math skill development.

6.2.3.3 Results and discussion

The model shown in Equation 6.6 is fit to data of each GLOP separately in the statistics software package R (see www.r-project.org). To account for variance among students and items, student IDs and item IDs are also introduced as factors. By taking this step we take into consideration the fact that student responses are not independent of each other, and properly compute statistical reliability and standard errors. After the model is fitted, it outputs estimated coefficients for every item in each GLOP. Table 6.3 reports the estimated value of the *B* parameters of items in two GLOPs, GLOP 1 and GLOP 4, and also the standard error, order descending by the value of B in each GLOP. We can see that among

all the items in GLOP 1, Item 1022 has the largest positive impact on student skill development: 0.464 in scale of *logit* (although the logit (log-odds) scale is not the most common one, it has the property that the item with the largest B coefficient will result in the largest learning gain, and an increase of 0.464 in logit scale is approximately equivalent to an increase of 0.116 in the probability of giving a correct answer). Unfortunately, the model has determined that working on Item 1649 does not help student learning, indicated by a negative value of B although the value is not reliably lower than zero.

Table 6.3. Coefficients of logistic regression model for items in GLOP 1 and GLOP 4

GLOP ID	Item ID	B (Coefficient) (higher is better)	std. err
1	1022	0.464	0.257
1	1660	0.414	0.247
1	1045	0.127	0.254
1	1263	0.011	0.241
1	1649	-0.176	0.261
4	2264	0.707	0.225
4	2236	0.079	0.236
4	9086	-0.014	0.232
4	2239	-0.236	0.237
4	2274	-0.274	0.240

p =.08

p =.1

p=.05

It looks like the items vary in their instructional effectiveness in helping student learn the skill(s) associated with a GLOP. But the standard errors are relatively large, too. So, given any pair of item *i* and item *j*, I perform z-test to determine whether coefficients for item *i* and item *j* in the logistic regression model are statistically significantly different ($p<0.05$). The z-score is calculated using $z = (B_i - B_j) / \sqrt{stderr_i^2 + stderr_j^2}$, assuming a normal distribution. As shown next to Table 6.3, the model fails to reliably tell the difference among the top 4

items in GLOP 1 (# of items = 5; # of students = 531; # of data points = 2,256), but only find marginal difference ($0.05 < p < 0.1$) between Item 1022 and item 1649, and between Item 1660 and Item 1649. However, we succeed to detect difference between the top 2 items in GLOP 4, Item 2264 and Item 2236 ($p = 0.05$) (# students = 652; #data points = 2,573).

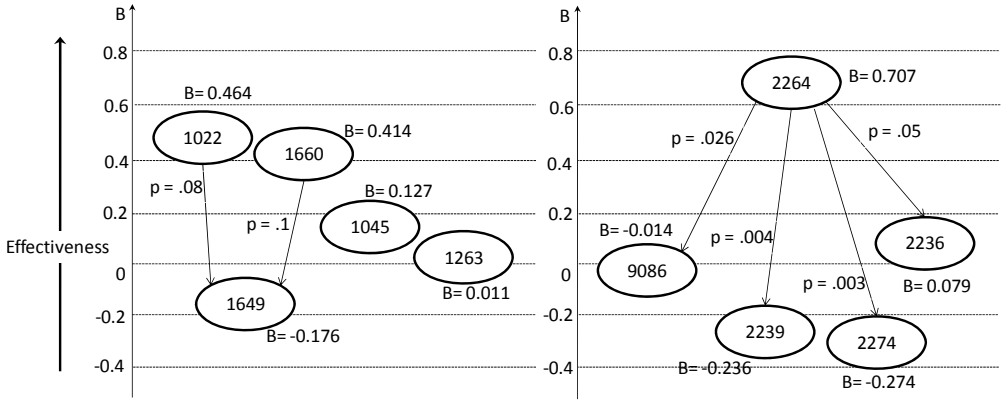


Figure 6.2. (a) Partial order relationship of items in GLOP 1; (b). Partial order relationship of items in GLOP 4.

Ideally, I would like to come up with a partial order of items in a GLOP that reflects which item caused the most learning, which one comes next, and which item is least effective. Figure 6.2(a) illustrates the partial order relationship among items in GLOP 1. In the diagram, an arrow connecting two items suggests there is a reliable difference between the instruction impact of the two items; and the arrow points to the less effective item. Additionally, the higher an item locates in the diagram, the larger the estimated value of B is. We follow the same process to

acquire orders of items in all 39 GLOPs in our data set. The partial order diagram for items in GLOP 4 is shown in Figure 6.2 (b).

One question is what types of items lead to more learning, easier or harder ones? Presumably, I hypothesize if a student learned to solve a hard item, he/she then should be able to do better on an easier item that requires similar skills. However, the converse is not necessarily true. In the last section, I tested this hypothesis and our results indicate that students learn as much from easier items as from harder items. Thus, those results suggest rejecting the hypothesis. Here, I will replicate the investigation: we have directly estimated the amount of learning caused by each item (B parameter), and also since item ID was used as a factor in the logistic regression model, we get one parameter each item that reflects the easiness of the item. Interestingly, this time we are able to find a significant correlation between the two values (the amount of learning vs. easiness) of the 181 items (Spearman's rho = .192, p = .010), which suggests that, although the effect is not large, in general students did learn more by doing easier items. This result makes sense from the perspective of cognitive development. A hypothesis proposed by a cognitive scientist, Kenneth Koedinger (personal communication), was that easy questions were easy because they only required the usage of a single skill, or fewer closely related skills. Hard questions were hard because they involved multiple (or extra) skills. The extra required skills were not intrinsic to the GLOP, and thus practice on them should not be helpful on other items in the

GLOP. But easy items forced students to focus on the crucial part of the GLOP. Thus, practicing these items helped students to perform better later in the same GLOP.

Above I explore the research question of measuring the instructional effectiveness of different problems, and associated tutoring, using the learning decomposition technique. One area of interest following this work is how to validate these numbers. One approach is to use human raters. While doing the learning decomposition analysis, we invite two human experts to review the items in 25 GLOPs that include less than 4 items. The human experts are asked to come up with a ranking of the items based on which items (including the scaffolding questions and the hint messages) they think will produce the most learning. It took the experts about 3 hours to finish ranking all 68 items. I then look at the pair-wise correlation between rankings of the two human experts, and the ranking rendered by the learning decomposition approach. It turns out that the rankings of the experts correlate with each other significantly (Spearman's rho = .238, p = .049). Yet, I can only find reliable correlation between one rater's ranking and the learning decomposition ranking (Spearman's rho = .323, p = .007). Notice that this correlation is even stronger than that of the two human raters, which provides some evidence for the validity of our results. However, overall, the inter-rater reliability is relatively low, so we will need to try harder to obtain stronger evidence. Another approach is to make use of synthetic data by using a computer

simulation study. The benefit of using a simulation is that we would know the ground truth about the effectiveness of each item. Furthermore, running a simulation study would allow us to better understand the power of the learning decomposition approach, such as how big the differences between the learning impacts of two items need to be for this approach to be able to detect it? How many students, data points the learning decomposition approach requires to tell a given difference between two items?

To summarize the section, I have shown how learning decomposition can be applied in the domain of mathematics to use observational data to estimate the effectiveness of different tutoring content. It provides an evidence that the learning decomposition is not domain specific and generally applicable to a variety of ITS that focus on different domains. An open issue for this approach is related to the generalization of the approach. There are other factors that we have not yet explored such as the variant item effectiveness for students of different knowledge levels. Applying the methodology in other domains, esp. ill-defined domains, possibly involves analyzing more other factors. The results may be affected by the organization of GLOPs as well.

6.3 Towards refining existing skills models

Existing literature has shown that creating an accurate model of a students' knowledge can be quite difficult due to various sources of uncertainty caused by factors such as multiple sources of student errors, varied problem solving strategies, problems associated with more than one skill, careless slip and lucky guesses, learning and forgetting (Katz, Lesgold, Eggan & Gordin, 1992), requiring the time of experts to create and then test these models on students. The expert-built models are subject to the risk of "expert blind spot" (Nathan, & Koedinger, 2000). I am happy to see that our first cognitive model fits relatively well on student performance data. Nevertheless, it is not an optimal model, esp. considering the amount of time we spent on developing the model, and this first model is just our "best guess". I still feel that we can probably refine the fine-grained model to be more accurate. For instance, as I have noticed from the modelling results I get from Chapter 5, for DATA-2005, among the 78 skills in the WPI-78 model, 13 coefficients indicate un-learning (i.e. the slopes are negative). This is a sign that the tagging of problems with these 13 skills might be appropriate and can be further improved.

6.3.1 Using data mining findings to aid manual update of existing models

So, given the problems, how do we refine the existing model? The first approach I can think of is to hand the model back to our subject-matter expert and ask her to improve the model. However, as mentioned before, our existing fine grained skill model is "hand-crafted" and involves more than 1,000 items. Therefore, it is not plausible for the content expert to examine every item to improve the model, especially if we need to do this iteratively and want this to be done in a timely manner. Well, here is where the data analysis results can provide some clues. One clue is to focus on the "un-learned" skills. One reason that a skill might have a poorly fit slope would be that we tagged items with the same skill names that share some superficial similarity, but do not have the same learning rates. As mentioned above, there are 13 of these skills whose tagging should be further examined. In terms of items, the candidates to be examined would be items for which the mixed-effects model produces highest residuals, and those for which student performance has been consistently over-predicted or under-predicted across all students.

A second clue comes from my item level analysis of instructional effectiveness as described in the last section. In that work, the first finding is that the amount of learning is rather uneven across GLOPs. Remember GLOP is a group of items that are organized into one group by subject matter experts

because they are associated with the same skill in the skill model. So, the GLOPs from which students did not show much learning raise a signal that maybe these items do not belong to the same group, so there was no transfer among them. Finding out the GLOPs that did not causing learning suggests potential problematic tagging with the group of items in the skill model. While when we take an even closer look at items in each GLOP, I notice which bits of content were not very effective in promoting learning in ASSISTments system. One reason that these items produced least learning could be the quality of tutorial materials of the items is relative low. But these items are also candidates for which the skill-tagging shall be examined.

In this way, we use automated data mining methods to help intelligent tutoring system builders learn from the logs of students data what portion of their tutors worked well and what portions should be improved. And then, this information can be used to aid content experts so that they can concentrate on the most problematic skill tagging and gain efficiency on model improvements. Anecdotally, our subject-matter expert is happy about the results we present to her and think they are helpful. But, at the moment, she is still in progress of updating the existing skill model and we do not yet have empirical results to quantitatively demonstrate whether the data analysis improve efficiency substantially.

6.3.2 Searching for better models automatically

The first approach of model refinement I describe in the last section depends on human intelligence and experience, while the second one relies on computational capability of machines by improving and applying a method called learning factor analysis.

6.3.2.1 Description of learning factor analysis

Cen, Koedinger & Junker (2005, 2006) proposed a generic method called learning factor analysis (LFA) for cognitive model evaluation and refinement. LFA was initially conceptualized by Koedinger and Junker (1999). It aims to "combine statistics, human expertise and combinatorial search to evaluate and improve a cognitive model". LFA has three parts: a statistical model that evaluate how the model fit the data; difficulty factors associated with problems; and a combinatorial search space of cognitive models to for student models that better account for student performance data directed by the evaluation of the statistical model.

LFA is an extension of the power law of learning (Newell & Rosenbloom, 1993), which represents that the error rate decrease exponentially according to a power function as the number of opportunities to practice a skill increase. The power law applies to one particular student and over only one skill while LFA models multiple students and multiple skills by adding in student and skill

intercepts and skill learning rates. Cen, Koedinger & Junker (2006) used a multiple logistic regression model:

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

where

p = the probability to get an item right

X = the factors for students

Y = the factors for skills

T = the covariates for the number of opportunities practiced on the skills

Y T = the covariates for interaction between skill and the number of practice opportunities for that skill

α = the coefficient for each student, i.e. the student's "smarts"

β = the coefficient for each rule, i.e. the skill's difficulty

γ = the coefficient for the interaction between a skill and its opportunities, i.e. the learning rate.

In LFA, a factor is a hidden feature in a problem that makes the problem easier or harder to solve. It is usually identified by subject exerts based up instruction theory and task analysis. An example factor in math with two possible values is using a rule CIRCLE-AREA (e.g. $S = \pi * r^2$) *forward* (to calculate circle

area given radius) or *backward* (to calculate radius given circle area). Here *forward* and *backward* are values of a difficulty factor.

LFA performs heuristic search over a search space where each state is a new cognitive model to locate the best one. Starting from a base model, it applies one of the 3 operators "split", "add", and "merge" on skills in current model to generate sub-models by incorporating the difficulty factor (or factor). For instance, operator "add" means that the factor with its possible value is simply added as a new skill to the original model. Suppose solving a problem requires applying CIRCLE-AREA backward. By applying the "add" operator, there will be one more skill called "APPLYING-FORMULA-BACKWARD" added to current model and tagged to the problem, while the rest skills remain unchanged. In this way, a sub-model is generated and the search space expanded. Various heuristics such as AIC, BIC, R-square and Log likelihood, have been considered as model evaluation and selection measures.

6.3.2.2 Suggesting possible factors for LFA search

As a basis of LFA, difficulty factors have always been found by subject experts through a process of "difficulty factor assessment" (DFA) (Koedinger, 2000). Based upon theory or task analysis, researchers can hypothesize the likely factors that cause student difficulties, and by assessing performance difference on pairs of problems that vary by only one factor, the experts identify the hidden

knowledge component that can be used to improve a skill model. It is phase of human making hypothesis and identification that makes LFA a semi-automatic approach, even though it is intuitively appealing as being a fully automated method.

Can we raise efficiency of LFA by suggesting difficulty factors automatically yet still get better models? In Section 6.2, I have shown certain items in a random sequence cause significantly less learning than others. Intuitively, it is highly possible that there is certain factor inherited in the items, which make it harder for the learning from this item to transfer to later items. This could be either because later items demand more skills than the current one, or because what a student learns from current item does not help later items. In both conditions, there is probably "mis-tagging" with this item. Thus, presumably, such a factor can be utilized by LFA to manipulate the original skill model to search for the best-fit model. What I really hope to see is that having a human expert sitting in front of a computer, with the help of our educational data mining results, she can quickly determine what factors each item may have. Before doing that, I want to check to see if my results can be used to make suggestions on factors and whether there is some validity in the approach.

In order to test this idea, I create factor tables for all the GLOPs as described in Section 6.2.1. In each table, I use one factor with two values "High" and "Low"

indicating the effectiveness of the items. The item that has caused least learning is associated with "Low" while all other items are associated with "High". Table 6.4 shows the factor table for GLOP 1. Noticing that since all items are tagged with the same skill "Interpreting-Circle-graph" as they all belong to one GLOP.

Table 6.4. Assigning factor to GLOP 1 based on learning coefficients

GLOP ID	Item ID	Skill	Factor	B (Coefficient) (higher is better)
1	1022	Circle-graph	High	0.464
1	1660	Circle-graph	High	0.414
1	1045	Circle-graph	High	0.127
1	1263	Circle-graph	High	0.011
1	1649	Circle-graph	Low	-0.176

6.3.2.3 Results

Given the factor tables, I run LFA search over all the GLOPs. Following Cen, Koedinger & Junker (2005, 2006), BIC is used as the heuristic to evaluate the models in that it balances simplicity and predictive power of models. Among the 38 GLOPs I have examined, LFA is able to find statistically significantly better models (a difference of 10 points or more on BIC) for 12 of them, using the factors as assigned in the factor tables. Among the 12 GLOPs, 5 of them include 2 items; 3 include 4 items; the rest 4 GLOPs have 5, 6, 8, and 9 items respectively.

To get a feeling of how well the suggested factors do, I conduct a simple sanity check where I randomly assign one item each GLOP with the "Low" value of the factor, and then run the same searching process as before. Obviously, for the 2-item GLOPs, the results will be the same as before. But for GLOPs with

more items, the search process using randomly assigned factor values only find better models for 2 out of the 27 GLOPs, which makes our previous results of 7 out of 27 somewhat impressive.

I admit there are other ways of assigning values in the factor table. Yet, I am also glad to see that the results show some validity for the very simple way of suggesting factors. I am especially happy to see that when there are more than 5 items in a GLOP, this method can still help find better models for 4 out of the 15 GLOPs.

This work is preliminary despite of the inspiring results, in that the amount of data I have applied this method to is very limited. I would like to apply this approach on data collected from other tutoring systems to verify the generality as well. Moreover, I do realize that using human experts' suggested factors would be another control condition to compare to. But considering the amount of efforts and time that need to be spent on difficulty factor assessment, maybe it would not be a totally "fair" comparison. Another reasonable study would be to run a randomized controlled study to compare two conditions where in one condition, human experts use solely DFA to identify factors, while in the other condition, human experts are provided the item level tutoring effectiveness results as I show in Section 2. The study should be controlled for time, and then controlled for groups to examine on what aspects the results can be helpful.

6.4 Constructing skill models for different groups

6.4.1 Introduction

The topic of representing domain knowledge is fundamental in the construction of intelligent tutoring systems. This representation is important not only because it denotes the language used in constructing the tutor (e.g. the level at which to construct hints), but also because it makes claims about the level at which students represent knowledge and transfer it between problems. For this reason, such models are sometimes called transfer models (e.g. Croteau, Heffernan, & Koedinger, 2004).

In the last section, I have shown how our results on item instructional effectives can suggest "factors" for the searching process of LFA, and thus, improve model fit. While in this section, rather than trying to invent another complex and computationally intensive technique, I take an alternative view. We know from prior research that students of differing proficiency have somewhat different representations of the domain, with more skilled learners having a more compact (i.e. coarser) representation (Leszczenski & Beck, 2007); Rafferty & Yudelson, 2007). We also know that different tutorial interventions influence the representation that learners acquire, with better interventions causing learners to develop a more compact representation (Koedinger & Mathan, 2004). A common

fallacy is the belief that finer-grain models will fit learner data better, or at least will fit better given sufficient training data, since they are able to represent subtler distinctions in the domain. This belief is incorrect since fine-grained models not only make subtle distinctions in skills, they (typically) also assert that tasks are more independent of each other (being tagged with different skills). Thus, knowledge transfer happens less often in that practice in one task does not help with another. If learners are able to transfer knowledge amongst skills, a coarser-grain model will better fit performance data. Theoretically based upon the learning curve equations (Newell & Rosenbloom, 1993), a compact model allows for faster mastery of skills if learning rates do not decrease sufficiently to compensate for the greater number of practice attempts for each skill. Intuitively, consider a model with 2 skills and a student who needs five practice attempts for each skill to achieve mastery. If the model can be compounded into a more compact model of just 1 skill by merging two of the original skills, then literally the student will have ten practice opportunities on the new skill as every opportunity on either of the two merged skills will count as an opportunity on the new skill. Suppose the students' learning rate and the initial knowledge does not change a lot after the "merging", the student may not need all the ten practices to gain mastery, which means the learning speed is increased.

Given the previous results and the fact that we have built a series of hand-constructed transfer models at various levels of granularities, perhaps it makes

more sense to skip over automated techniques and simply start with existing transfer models and use them as a lens to examine student learning. In this way, we can still do interesting science with our large datasets but do not have to focus on complex machinery. Specifically, we are interested in whether models of different granularity better fit distinct subgroups, and, consequently, whether we can use this approach to evaluate schools by examining the grain-size at which their students are best represented. Given two schools where one is better predicted by a coarser transfer model, that school is probably the better one. This approach is different than simply looking at which school has the highest test score performance. If a weaker school changes its curriculum and its students have a better mental model of the domain and are transferring better, they might still lag a stronger school in raw knowledge and consequently in test scores. This approach can potentially detect such schools. We can validate this hypothesis in two ways. First, we have an idea of the quality of the schools we are evaluating (although the person interpreting the data did not). Second, instead of partitioning students by school, we can use their state assessment test score and partition them by math proficiency. If we see a trend for stronger students, it is reasonable to believe it applies to stronger schools.

The advantage of this approach is that it is easy. Also, if one transfer model does a better job at a particular school, since that model is expert-constructed it should not be (any more) difficult (than usual) to construct tutorial content for the

model; whereas automated models might not fit educators' understanding of the domain. For this research, I use data collected in the ASSISTments system as the test bed.

6.4.2 Approach

DATA-2005 (after preprocessing using scaffold-blame and partial credit strategies) as described in Chapter 5 is reused for this analysis. The only difference is that all 495 students in the initial data sets are used here. This changed the statistics slightly. The data set contains 147,624 data points from the 495 students, among which 45,135 come from original questions. The students have used the ASSISTment system on at least 6 days, with an average of 9 days during 2004-2005 school year and on average, finished 91 original questions.

The first portion of this research involves partitioning students into groups to determine if different groups of students have different patterns for learning math skills. Naturally, the 495 students can be separated by the schools they were in, with 312 from school F and 183 from school W. I also try to separate them by their performance level at the 2005 MCAS test. The high performing group includes the 128 students whose performance level is assessed by the state as "Advanced" or "Proficient"; the medium group includes the 154 students whose performance level is "Needs Improvement", and the low performing group has the rest 213 students at performance level "Warning". While these performance levels

are somewhat specific to Massachusetts, they are at least criterion-referenced and much more general than numbers extracted from a student model or raw scores on a test (what qualifies as "Proficient" in Massachusetts is probably similar to "Proficient" in Macedonia). The hypothesis is that students from a stronger school, or higher performing group, would show more transfer in their knowledge acquisition than those from a weaker school, or lower performing groups. Hence, for the stronger students and schools the coarser grained model will better describe their learning and provide more accurate prediction of their MCAS test scores.

Similarly as in Chapter 5, mixed-effects logistic models are fit here. But we have chose two prediction evaluating functions for this analysis, namely, mean absolute difference (MAD), and mean difference (MD), as below.

$$MD = \frac{1}{n}\sum_{i=1}^{n}(prediction_i - MCAS_i)$$

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

6.4.3 Results and discussion

The results for both school F and school W are summarized in Table 6.5. As shown in Table 6.5, school F has a flat error line across all four different transfer

models. The MAD for the WPI-39 model is the lowest, and yet a paired t-test that compares the absolute pair-wise differences of individual students among all models suggested that there is no reliable difference. However, for school W, the line tilts: the MAD of the WPI-39 model is reliably lower than those of the WPI-1 and WPI-5 models, indicating school W is better predicted by a finer grained model than by coarser grained models. Note that we are not able to fit the statistical model for school W with the WPI-106 transfer model (there is a technical glitch we do not understand and are investigating). We encounter the same problem later in the paper, which admittedly bring up some caveats in interpreting our results. The second part of Table 6.5 shows the values of MD for each model. The results indicate that both schools are optimized at the WPI-39 model. In general, student performance on the state test is overestimated by our models except that the WPI-106 model underestimates school F; and school W is even more overestimated than school F across known results from all the three models. As we know that, theoretically a one-skill model assumes perfect transfer. Since that is unlikely to happen, it would tend to overestimate student performance. And for a weaker school, perfect transfer is even more improbable. Thus, the overestimation would be greater since students are probably learning a collection of 106 unrelated skills. The tendency of overestimate decreases as the granularity of transfer models increases, and a very fine grained model such as the WPI-106 model that assumes no transfer or very low transfer may even

underestimate when there is actually some level of knowledge transfer. We can see that in Table 6.5, the MD goes from positive to negative when we use the WPI-106 model for School F. Given these results, based on our hypothesis we would predict school F is the stronger school. An examination of both schools' MCAS performance reports (for current achievement) and information on their Annual Year Progress (AYP, for changes in performance) confirms our prediction.

Table 6.5. Results for students grouped by schools

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

Table 6.6. Results for students grouped by performance levels

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

As mentioned in section 6.4.1, a second validation approach is that instead of partitioning students by school, we can use their state assessment test score and partition them by math proficiency. If we see a trend for stronger students, it is reasonable to believe it applies to stronger schools. As discussed in section 6.4.2, I split all the 495 students into 3 groups based on their state test performance level, and fit a mixed-effects logistic regression model to each group separately for different transfer models. The values of MAD and MD are summarized in Table

6.6. I see a slight support with MAD: for the students at the high end, the WPI-39 does the best job at predicting their state test scores, reliably better than the other three models, while the WPI-106 model does reliably worse than the WPI-1 and WPI-5 models, suggesting there is certain amount of knowledge transfer happening with the high performing students. However, since we do not obtain results of the WPI-106 model for the other two groups, it is hard to draw a conclusion there. When it comes to the MD measure, we notice some support as well. Obviously, the advanced and proficient students have been underestimated by all models, and the amount of underestimation goes worst when the finest grained model, the WPI-106 model, is applied. On the contrary, the medium and low performing students are all overestimated under all the models. Just as we hypothesize, the finer grained models overestimate less than the coarser grained models, and the better performing, stronger groups are less overestimated than the weaker groups. Therefore, weaker students are better represented by transfer models that are finer-grained.

6.4.4 A bottom-up aggregation approach

Rather than starting with an *a* priori disaggregation, I also try a bottom-up aggregation approach, focusing on treating students as individuals and discovering commonalities among students who are best-fit with a particular transfer model. We have collected demographic data about several properties of a

student, such as which school he/she goes to, ethnicity, gender, etc. Finding out the relation among these properties and which transfer model best fits this student is the goal. The plan is to bring together model-fitting information and student characteristics, and then use a machine learning classifier to determine the best-fit model. This bottom-up aggregation is a strong alternative to proposing and testing disaggregation, and will scale nicely as we get more descriptors for each student.

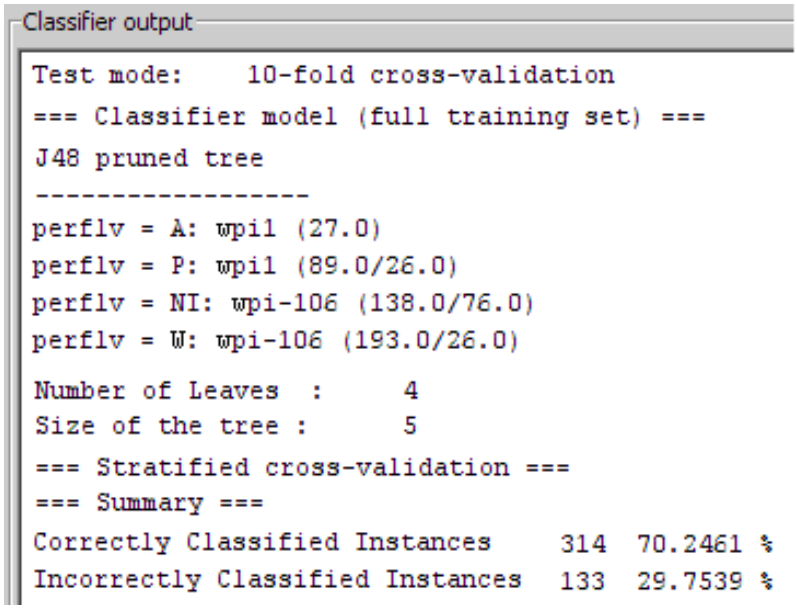


Figure 6.3. Result of classifying in Weka using J48 pruned tree

For this purpose, I first re-fit models for all the students as one group[33] and identify which model best fits each individual student. The best-fit model information is then combined with other properties of the student in a new data set.

[33] Since we are fitting models for all students at the same time, I have to use the data set of only 447 students simply because of a memory limit of R.

Specifically, the properties we use are: gender, free-lunch status (indicative of family income), special education status, ethnicity, and state test performance level. These properties are picked because they are easy to access, and all of them have meanings to researchers working with other populations in other locations. In comparison, properties such as the school a student attends are much less useful to those in other locations. Given the new data set, we built a J48 (C4.5 revision 8) decision tree in Weka 3.6. The constructed J48 pruned tree is show in Figure 6.3 that tells how the classifier uses the attributes to make a decision. The constructed tree is extremely simple with just 5 nodes. The WPI-1 model is overall the best fitting model for Advanced (A) and Proficient (P) students, and the WPI-106 is for "Needs improvement" (NI) or Warning (W) level students. The numbers in brackets after the leaf nodes indicate the number of instances assigned to that node, followed by how many of those instances is incorrectly classified as a result. In our case, the correct classification rates are relatively good for students at performance level of A, P, and W. Yet, for students at performance level of NI, even though the WPI-106 model is the best fit, it is not dominant with 76 out of 138 instances misclassified. It is encouraging that this simple decision tree can achieve a predictive accuracy of over 70% during stratified cross-validation. Although the decision tree only uses MCAS performance, it is provided with the variables described above but is unable to find a use for them. This result suggests the appropriate level of transfer model

granularity really seems to depend on student knowledge, rather than on variables that may correlate with knowledge such as family wealth. Therefore, if tutor designers have students with rather different levels of knowledge, they might wish to use different levels of their skill hierarchy. This point does not contradict the use of evaluating interventions (Koedinger & Mathan, 2004) and schools by model granularity: other properties certainly matter in how well knowledge transfers, but for our dataset they are not as predictive as the student's knowledge.

In summary, in this section, I start with existing hand-constructed transfer models at various levels of granularity, and use them as a lens to examine student learning. Specifically, I start by examining whether we can evaluate schools by determining the grain-size at which its students are best represented. I have also examined what models best fit students at different levels of proficiency, and found some support for the idea of stronger students being better fit with coarser transfer models. The interesting analysis is the bottom-up aggregation and using classification to find clusters of students who learn similarly. Through the usage of a bottom-up aggregation approach, the problem is changed. Rather than trying to automate the model search, why don't we automate seeing which student best fits which model? Results of the analysis suggest transfer model granularity really seems to be about student knowledge.

6.5 Related Work and Contribution

My work on using educational data mining approach to analyze relative instructional effectiveness of individual tutoring contents in intelligent tutoring systems is among the earliest efforts on this topic. Pardos & Heffernan (2009) explored the same question using the same data set but a different, Bayesian network approach. Even though the Bayesian networks is generally considered powerful, the learning decomposition method is an easier recipe, and also more scalable when the number of items increases. The traditional approach uses randomized controlled studies that can be very expensive, speaking of the time and subjects, especially if we want to focus on hundreds of individual pieces of tutoring contents. Yet, I demonstrate an approach where we can apply educational data mining technique to use observational data to efficiently compare the effectiveness of individual pieces of instructional materials in a tutoring system. This is a low cost approach comparing to randomized experiments.

Given the importance of transfer models, it is not surprising that their construction and improvement has been a major focus in the community. At the end of Chapter 5, I have reviewed many works in the literature on constructing transfer models. Beyond that, Winters et al. (2005) has compared a variety of statistical approaches for constructing transfer models, including cluster methods such as k-means and dimensionality reduction such as non-negative matrix

factorization. One common thread of this work is that it produces models that are typically more compact than those created by experts. This difference is both a source of strength (perhaps students learn differently than experts believe?) and a source of weakness (if the models are less understandable or make it harder to represent pedagogical knowledge why should we use them?). Although it would be an expensive undertaking, I am unaware of a controlled study showing that a tutor using automatically constructed model provides superior teaching compared to a tutor using to hand-constructed transfer model (or vice versa).

Rather than inferring a transfer model from scratch, LFA is a hybrid approach of refining existing models. As I have described in Section 6.4.2, this technique starts with a transfer model, typically built by hand, and computationally tries various modifications to the model to better align it with student performance data (e.g. see Cen, Koedinger & Junker, 2005, 2006). LFA has also been used in Leszczenski & Beck (2007) and Rafferty & Yudelson (2007) to answer scientific question on the level of knowledge that learners use to represent written words and problem-solving skills in geometry. Pavlik, Cen, & Koedinger (2009b) reconfigured LFA by taking into the desired features of knowledge tracing (Corbett & Anderson, 1995), and proposed a new version called performance factor analysis (PFA). Different from LFA, PFA uses two parameters to track a student's prior successes for a skill separately from the student's prior failures for the skill.

Different from LFA that uses computation intensive method to search for better models, Ferguson, Woolf, & Mahadevan (2009) developed a method to use transfer learning to guide the improvement of skill models. They hand-coded the transfer features in problems and thus constructed a hierarchical transfer learning model as an improvement of existing flat skill model.

Ritter et al. (2009) addressed the issue of model improvement by investigating the nature of the space of parameters from knowledge tracing. They used a k-means clustering to drastically reduce the parameter space used to model students from 2,400 skills to 23 clusters without compromising the behavior of the system (the Cognitive Tutors).

6.6 Conclusion

In this chapter, I first present results suggesting that students are learning from working within ASSISTments and also point out that the learning is rather uneven across skills. I then take an even closer look at individual items that are tagged with the same skill and find out that the instructional effectiveness varies across items. It is found out that some items cause significant learning while some other items are not as useful at promoting learning. This is among the first efforts to look at effectiveness of individual item. I argue that this is an important step towards improving overall effectiveness of a tutoring system.

Following my effort on detecting item effectiveness, several practices are described on how to use the results to improve skill modeling in ASSISTments. Possibilities of both manual approach and automatic approach, i.e. learning factor analysis, are considered. Preliminary results show our findings in item effectiveness from educational data mining can be used to assign factors, and thus, automate LFA. As another way to improve the predictive power of skill models, the hand-constructed models of various granularities are revisited to explore the correlation between student knowledge and the grain-size of their best-fit skill models. I find out that high performing students are better modeled by coarser grained models while finer grained models are more predictive for lower performing students.

I want to acknowledge that we don't have control group to compare the learning results against to. Also, if student performance systematically varies over time apart from learning, my model does not account for it. For instance, if students experienced a ramp up effect of doing better over time, this could explain away our results. Similarly, if students get fatigued over a class period we would be underestimate the learning effect.

Currently, our GLOPs are manually constructed. It would be interesting to see how the items would be grouped by some automated method such as Q-matrix

algorithm or LFA, and then how our results will be impacted by a new grouping method.

Chapter 7

7 Conclusion of this Dissertation

7.1 Conclusion

The ASSISTments system has been launched and is in its fifth full year of operation. Unlike other assessment system, it provides instructional assistance while assessing students. The dissertation focuses on how to assess students better in such a learning environment. In this dissertation, I address the testing challenge in this web-based learning and assessment system from different aspects.

Teachers' needs have been addressed since the very beginning of the ASSISTments project. An online reporting system is developed to inform teachers of their students' performance and where students are having difficulties.

In this dissertation, instead of ignoring students' learning behaviors as most traditional tests do, I leverage this information in our predictive model. Some evidence is presented that the online assessment system can do a better job of

predicting student knowledge by being able to take into consideration how much tutoring assistance was needed, how fast a student solves a problem and how many attempts were needed to finish a problem. Traditional assessment usually happens in a limited amount of time. Yet, in this dissertation, promising evidence is found that the ASSISTments system is able to track students' learning during a year well. Furthermore, our prediction of students' end-of-year test score is as accurate as the standardized test itself.

Being able to estimate student's total score is not enough. In this dissertation, I help build a fine grained skill model that is used to model student's individual piece of knowledge components separately. The results from rigorous evaluation show that the fine-grained model can estimate student proficiency more accurately than existing coarser-grained models. More importantly, such a model is used to provide teachers with useful fine-grained student-level knowledge they can reflect on and adjust their pedagogy.

Realizing the skill model is not optimal, I am also seeking ways to improve it. The application of educational data mining techniques helps find out the skills that students have a hard time mastering. Relative effectiveness of individual pieces of instructional contents can be detected as well. Since skills are "latent", its mastery is reflected by performance on the problems tagged with the skills. The results thereby identify potential targets for skill model improvements as well.

In this dissertation, both manual and automatic approaches are considered as ways to make use of the findings to refine existing models.

All in all, I believe that the ASSISTments System can be a better predictor of state test scores and fine grained knowledge because of this dissertations' work. With that said, I would like to end with a tantalizing question: Are we likely to see states move from a test that happens once a year, to an assessment tracking system that offers continuous assessment (Computer Research Association, 2005) every few weeks? While more research is warranted, my results suggest that perhaps the answer should be yes.

7.2 General Implications

A recent Computer Research Association (CRA) report (Computer Research Association, 2005) prepared for NSF reported on the need to fund the cyber-learning, and pointed out that continuous assessment systems research is a huge growth area. A general implication from this research suggests that continuous assessment systems are possible to build and that they can be more accurate at helping schools get information on their students. We argue that this result is important because it opens up the ability to blend assessment and assisting. This opens up a whole new area of assessment. To respond to Arne Duncan's call for better testing and data management systems, the studies from this dissertation

imply that it is possible for the states to develop such a system similar to the ASSISTments System that does all three of these things at the same time: 1) accurately and longitudinally assesses students, 2) gives fine grained feedback that is more cognitively diagnostic and 3) saves classroom instruction time by assessing students while they are getting tutoring.

Another idea upon which we can reflect is, "what is the right way to judge a project like this one, which tries to blend assessment and assisting (increasing student learning)?" The system does not have to be either the best assessment systems in the world, or the best learning system in the world. It needs to be a good balance between the two. Fundamentally, there will always be tradeoffs between the accuracy of assessment data and increases in students' learning, due to the fact that schools have only a finite number of days in a year.

Our results show that in the ASSISTments system, some contents have more impact on student math skill development while some contents are not so effective. I suspect that this result is not specific to ASSISTments, and other tutors have items that vary greatly in educational effectiveness. This study demonstrates another, low cost, approach of evaluating ITS contents other than experimental study. Potentially, the approach and the results can be used to examine the quality of instructional contents in a learning system, and thus improve the overall learning impact.

Automated techniques for revising transfer models for better knowledge representation have addressed interesting scientific questions. Is there a way we can do interesting science on educational data sets and avoid the "irritating" automation step? The answer from this dissertation is "yes," if it is possible to build a hierarchy of transfer models with different granularity. Previous experience tells us that hand-constructed transfer models at different grain sizes is not a rare thing to have, and not very hard to think about. The hierarchy can be used for runtime benefit of intelligent tutoring systems such as the control of mastery learning or generation of feedback messages for students of various proficiency levels. It can also be used to evaluate schools and be validated via high stake test performance. I argue that hand-created transfer models and a bottom-up approach to aggregating students is a better use of human brains and computational power than approaches that focus search efforts on revising the domain model. Better understanding what parts of the scientific enterprise can be best done by people and which are better done computationally is a major issue in the area of educational data mining.

7.3 Ideas of Future Work

7.3.1 Evaluating the impact of reports on decision making

In the ASSISTments project, we don't yet have a solid answer to the research question "Can teachers use the advice of this system to lead to higher student learning?" Conducting year long studies to see if students in classrooms with teachers that are using the system lead to higher MCAS gains than in classrooms where the teacher did not use the system would help answer this question. There are some preliminary results showing ASSISTments is effective. Yet, we are not sure how much has to do with the feedbacks we provide in the reports (or it is just more practice in ASSISTments). Thus, one of the future work is to have a better understanding if detailed reporting to teachers of students' difficulties can lead teachers to do more data-driven decision making. We collected self reports from teachers but they seem too rosy and we think that are trying to trying please the researchers at WPI. We could send in an army of classroom observers is try to divine how teachers are changing their instruction in response to this data, but this is not cost-effective or feasible. Instead one way is to look to see which teachers and when they seem to be looking at the data by logging teachers' usage of the reports (that we currently have not done yet). This can be done by emailing teachers a link to get their report but in order to see it they will have to click on a

link, which will allow us to tell which teachers are even opening the email (with informed consent of course). Of course teachers will also be able to log into their accounts and access the same information so that usage can be logged as well. And given the fact that these reports will contain hyperlinks to dig further into the data, we will be able to distinguish between the casual teachers that only open their report from a teacher that explores their data in greater detail. Do the teachers continue to look at the data all year? If their self-reports correlate with the computer accesses then they are more likely to be believable. Are teachers that access the data more often the same teachers that are getting higher gain scores? Of course this is only correlation data. But running a randomized controlled experiment giving the system to some teachers but not others could provide some causal evidence.

7.3.2 Helping teachers to change their practice

This dissertation has been focusing on student modeling and better assessment. Yet, I want to address that student performance modeling and assessment is not the ultimate goal, but just gives the roadmap for change. Early findings of the national *Study of Education Data Systems and Decision Making* indicate that although there is dramatic increase in teacher access to student data systems, the data from these systems are having little effect on teachers' daily instructional decisions (Means, Padilla, DeBarger, & Bakia, 2009).

Recently in the speech, Arne Duncan, the United States Secretary of Education, called for teachers to use data to drive instruction in order to improve student achievement. He has been critical of the fact that teachers are currently learning to use data to drive instruction *on the job*, instead of that being part of their teacher preparation program and in service. He pointed out that "one of our collective challenges is to talk about data and research and ways that people understand."[34] As a matter of fact, a critique for our reports is that our feedback might be useful in identifying what skills that are used in the missed items are the ones that are causing the most trouble, but this might suffer from the fact that they are abstract and then hard for a teacher to know how best to respond. Thus, instead of overwhelming teachers by reporting on a lot of skills, maybe it is better to give advice of skills that are the "low hanging fruit" for their students by focusing on concepts and skills that are most directly to be within an individual student's reach. Another way to address this concern is to direct teachers to lesson plans and other activities. In education research area, researchers have recently started some practices on aiding this change on top of presenting them assessment results. For instance, the Diagnostic Geometry Assessment (Russell & Masters, 2009), an online assessment, provides a collection of lesson plans and instructional resources targeting common misconceptions. Teachers also have

[34] http://www.ed.gov/news/speeches/2009/06/06082009.html

access to follow-up tests that can be administered to the student after instruction to measure whether the student has corrected the misconception. At WPI, we have a partnership program in Math and Science education (PIMSE[35]) that involves participants of teachers, graduate students, and ASSISTments coach. The goal of the project is to help teachers figure out the best ways of integrating data-driven instruction into their own classroom practices. A workshop has been set up for 17 local teachers in the next school year and a college seminar entitled *using advanced educational technology to support data-driven decision making*[36] is available for teachers to take. Though it is a challenging task, I believe in that we just need to "being much more thoughtful about how we look at assessments and create incentives so that every child is pushed to excel and pushed to reach their potential" (said Arne Duncan).

7.3.3 Improving assessment work

The online, dynamic metrics have been shown to be very helpful at predicting student MCAS scores. Although effective, the metrics are simply counts and averages. A piece of future work following that could be to improve the online assistance metrics. For instance, since the number of hints available is different across problems and the amount of information released in each level of hint

[35] http://teacherwiki.assistment.org/PIMSE_Schedule_2009-2010

[36] http://nth.wpi.edu/MME562.htm

differs too, instead of simply summing-up or computing the mean value, we want to construct some weighting function to better measure the amount of assistance students requested to solve a problem.

Another piece of work follows up the assessment work is to predict fine grained knowledge across years. Since our model is clearly capturing something that is predictive of student future performance (as shown in Section 4.5.2), we are considering focusing on determining what predicts specific deficits in an area. The research question we want to answer will be: can an 8th grade student model be used to predict the student will have a problem with a specific 10th grade skill? Teachers will be glad to know the answer so that they can adjust their instruction to better help student knowledge learning.

7.3.4 Assessment for learning

Now that this dissertation shows that we can do a good job assessing students, it is reasonable to move to the next stages of model-based inquires (Mislevy, 2006) and think about how to use the assessments to make better, more adaptive intelligent tutoring systems. It would be rather a pity if the work stops on the power of a rich set of elements to predict performance on a standardized test, but never feeds back into the runtime tutoring process. Following this direction, it think future work could involve creating detailed learner models that based on the findings from this dissertation and then use the models inform teachers, parents,

learners and, more importantly, the tutoring system on the ways to best remedy problems. This way, we are not only doing assessment *of* learning, but also doing assessment *for* learning.

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

Bibliography

[1] Ahn, J., Brusilovsky, P., Grady, J., He, D., and Syn, S. Y. (2007) Open user profiles for adaptive news systems: help or harm? In: Proceedings of the 16th international conference on World Wide Web, WWW '07, Banff, Canada, May 8-12, 2007, ACM Press, pp. 11-20

[2] Aleven, V.A.W.M.M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, 26(2).

[3] Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, D. (2007). Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement,* 44(4), 341-359.

[4] Almond, R.G., Shute, V., Underwood, J. & Zapata-Rivera, D. (2009). Bayesian networks: A teacher's view. *International Journal of Approxmate Reasoning*. 50, 450-460.

[5] Anderson, J.R. (1993). *Rules of mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[6] Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[7] Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4 (2), 167-207.

[8] Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. LEA.

[9] Anozie, N.O.& Junker, B.W. (2006). *Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system.* American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA.

[10] Ayers, E. & Junker, B.W. (2006). *Do skills combine additively to predict task difficulty in eighth-grade mathematics?* American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA.

[11] Baffes, P., & Mooney, R. (1996). A novel application of theory refinement to student modeling. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*. Portland, OR. 1996. pp.403-408.

[12] Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. *Proceedings of 7th International Conference on Intelligent Tutoring Systems*. Maceio, Brazil.

[13] Baker, R.S., Roll, I., Corbett, A.T., Koedinger, K.R. (2005) Do Performance Goals Lead Students to Game the System? In *Proceedings of the 12th International Conference on Artificial Intelligence and Education*, 57-64.

[14] Barnes, T., (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*.

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

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

[17] Beck, J.E. (2006). Using learning decomposition to analyze student fluency development. *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems*. Jhongli, Taiwan. pp. 21-28.

[18] Beck, J. E., Chang, K.-m., Mostow, J., & Corbett, A. (2008). Does help help? Introducing the Bayesian Evaluation and Assessment methodology. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* Montreal, 383-394.

[19] Beck, J. E., Jia, P., & Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, 2, 61-81.

[20] Beck, J. E., & Sison, J. (2006). Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education*, 16, 129-143.

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

[22] Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5,7-74.

[23] Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2): 139-149.

[24] Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham, UK: Open University Press.

[25] Bock, R.D., Gibbons, R., & Muraki, E. J. (1988). Full Information Item Factor Analysis. *Applied Psychological Measurement*, 12, 261-280.

[26] Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9).

[27] Brown, A. L., Bryant, N.R., & Campione, J. C. (1983). Preschool children's learning and transfer of matrices problems: Potential for improvement. Paper presented at the Society for Research in Child Development meetings, Detroit.

[28] Brusilovsky, P. & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *Journal of Artificial Intelligence in Education*, 13(2-4):159-172, 2003.

[29] Brusilovsky, P., Sosnovsky, S., and Shcherbinina, O. (2005) User Modeling in a Distributed E-Learning Architecture. In: L. Ardissono, P. Brna and A. Mitrovic (eds.) *Proceedings of 10th International User Modeling Conference* (Edinburgh, UK, July 24-29, 2005). Lecture Notes in Artificial Intelligence, vol. 3538. Berlin: Springer Verlag, pp. 387-391.

[30] Brusilovsky, P. and Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In: P. Brusilovsky, A. Kobsa and W. Neidl (eds.): *The Adaptive Web: Methods and Strategies of Web Personalization*. Lecture Notes in Computer Science, Vol. 4321, Berlin Heidelberg New York: Springer-Verlag, pp. 3-53.

[31] Bull, S. & McEvoy, A.T. (2003). An Intelligent Learning Environment with an Open Learner Model for the Desktop PC and Pocket PC, in U. Hoppe, F. Verdejo & J. kay (eds), *Artificial Intelligence in Education*, IOS Press, Amsterdam, 389-391.

[32] Bull, S. & Nghiem, T. (2002). Helping Learners to Understand Themselves with a Learner Model Open to Students, Peers and Instructors, in P. Brna & V. Dimitrova (eds), *Proceedings of Workshop on Individual and Group Modelling Methods that Help Learners Understand Themselves, International Conference on Intelligent Tutoring Systems* 2002, 5-13.

[33] Bull, S. (2004). Supporting Learning with Open Learner Models, *Proceedings of 4th Hellenic Conference with International Participation: Information and Communication Technologies in Education*, Athens, Greece. Keynote.

[34] Bull, S. & Kay, J. (2007). Student Models that Invite the Learner In: The SMILI Open Learner Modelling Framework, *International Journal of Artificial Intelligence in Education* 17(2), 89-120.

[35] Bull, S., Gardner, P., Ahmad, N., Ting, J. & Clarke, B. (in press). Use and Trust of Simple Independent Open Learner Models to Support Learning Within and Across Courses. *User Modeling, Adaptation and Personalization* 2009, Springer-Verlag, Berlin Heidelberg.

[36] Campione, J.C., Brown, A.L., & Bryant, N.R. (1985). Individual differences in learning and memory. In R.J. Sternberg (Ed.). *Human abilities: An information-processing approach,* 103–126. New York: W.H. Freeman.

[37] Carmona1, C., Millán, E., Pérez-de-la-Cruz, J.L., Trella1, M. & Conejo, R. (2005). Introducing Prerequisite Relations in a Multi-layered Bayesian Student Model. In Ardissono, Brna & Mitroivc (Eds). The *10th International Conference on User Modeling 2005*.Springer. 347-356

[38] Cen. H., Koedinger K., & Junker B. (2005). Automating Cognitive Model Improvement by A*Search and Logistic Regression. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*.

[39] Cen, H., K. Koedinger, and B. Junker. (2006). *Learning factors analysis: a general method for cognitive model evaluation and improvement.* Presented at the Eighth International Conference on Intelligent Tutoring Systems (ITS 2006), Jhongli, Taiwan.

[40] Chang, K., Beck, J., Mostow, J., & Corbett, A. (2006, June 26-30). A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, 104-113.

[41] Clancey W. (1987). *Knowledge-based tutoring—–the GUIDON program*. Cambridge, MA: MIT Press; 1987.

[42] Collins, J., Greer, J., and Huang. S. (1996). Adaptive assessment of using granularity hierarchies and Bayesian nets. In *Proceedings of Intelligent Tutoring Systems*, pp. 569--577, 1996.

[43] Conati C., Gertner A., VanLehn K., Druzdzel M. (1997). On-Line Student Modeling for Coached Problem Solving Using Bayesian Networks. In

Jameson A., Paris C., Tasso C., (eds.), *Proceedings of the sixth International* Conference on User Modeling (UM'97). New York: Springer-Wien.

[44] Confrey, J., Valenzuela, A., & Ortiz, A. (2002). Recommendations to the Texas State Board of Education on the setting of the TAKS standards: a call to responsible action. from http://www.syrce.org/State_Board.htm

[45] Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

[46] Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995) Student modeling in the ACT programming tutor. Chapter 2 in P. Nichols, S. Chipman, & R. Brennan, eds., *Cognitively Diagnostic Assessment* Hillsdale, NJ: Erlbaum.

[47] Corbett, A.T. & Bhatnagar, A. (1997). Student modeling in the ACT Programming Tutor: Adjusting a procedural learning model with declarative knowledge. *User Modeling: Proceedings of the Sixth International Conference on User Modeling UM97 Chia Laguna, Sardinia, Italy*. New York: Springer-Verlag Wein. pp.243-254.

[48] Corbett, A. T., Koedinger, K. R.,& Hadley,W. H. (2001). Cognitive Tutors: From the research classroom to all classrooms. In Goodman, P. S. (Ed.) *Technology Enhanced Learning: Opportunities for Change.*Mahwah, NJ: Lawrence Erlbaum Associates.

[49] Computer Research Association. (2005). Cyberinfrastructure for Education and Learning for the Future: a Vision and Research Agenda. Final report of Cyberlearning Workshop Series workshops held Fall 2004 - Spring 2005 by the Computing Research Association and the International Society of the Learning Sciences. Retrieved from http://www.cra.org/reports/cyberinfrastructure.pdf on November 10th, 2006

[50] Croteau, E., Heffernan, N. T., & Koedinger, K. R. (2004). Why are algebra word problems difficult? Using tutorial log files and the power law of learning to select the best fitting cognitive model. *Proceedings of the 7th International Conference on Intelligent Tutoring System*. Maceio, Brazil.

[51] Daniel, B., Zapata-Rivera, D., McCalla, G. (2007) A Bayesian Belief Network Approach for Modeling Complex Domains. In A. Mittal, & A. Kassim (Eds.). *Bayesian Network Technologies: Applications and Graphical Models*. IRM Press. 13-41.

[52] Dean, T. & Kanazawa, K. (1989). A model for projection and action. *Proceedings of the 8th International Joint Conference on Artificial Intelligence*. pp. 985-990. Detroit: Morgan Kaufmann.

[53] Dimitrova, V., Self, J. & Brna, P. (2001). Applying Interactive Open Learner Models to Learning Technical Terminology, in M. Bauer, P.J. Gmytrasiewicz & J. Vassileva (eds), *User Modeling 2001: 8th International Conference*, Springer-Verlag, Berlin Heidelberg, 148-157.

[54] Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. Nichols, S. Chipman, & R. Brennan, eds., *Cognitively Diagnostic Assessment.* Hillsdale, NJ: Erlbaum.

[55] Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3), 495-515.

[56] Embretson, S. E. (1992). Structured Rasch models for measuring individual-difference in learning and change. *International Journal of Psychology*. 27(3-4):372-372.

[57] Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologist*. Hillsdale, NJ: Lawrence Elbaum Associate.

[58] Feldt, L.S. & Brennan, R.L. (1993). Reliability.. In R.L.Linn (Ed.), *Educational Measurement*, 3rd ed., pp 105-146. New York: American Council on Education and Macmillan.

[59] Feng, M., Heffernan, N.T, Mani, M., & Heffernan C. (2006). Using Mixed-Effects Modeling to Compare Different Grain-Sized Skill Models. In Beck, J., Aimeur, E., & Barnes, T. (Eds). *Educational Data Mining: Papers from the AAAI Workshop*. Menlo Park, CA: AAAI Press. pp. 57-66.

[60] Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006a). Addressing the testing challenge with a web based E-assessment system that tutors as it assesses. *Proceedings of the 15th Annual World Wide Web Conference.* ACM Press: New York.

[61] Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006b). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In Ikeda, Ashley & Chan (Eds.) *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems.* Springer-Verlag: Berlin. pp 31–40.

[62] Feng, M. & Heffernan, N. T. (2007a). Assessing Students' Performance: Item Difficulty Parameter vs. Skill Learning Tracking. Paper presented at the National Council on Educational Measurement 2007 Annual Conference, Chicago.

[63] Feng, M. & Heffernan, N. (2007b). Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in ASSISTment System. *Journal of Interactive Learning Research*. 18 (2), pp. 207-230. Chesapeake, VA: AACE.

[64] Feng, M., Beck, J,. Heffernan, N. & Koedinger, K. (2008). Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test? In Beck & Baker (Eds.). *Proceedings of the 1st International Conference on Education Data Mining.* Montreal, 2008.

[65] Feng, M., Heffernan, N., Beck, J, & Koedinger, K. (2008). Can we predict which groups of questions students will learn from? In Beck & Baker (Eds.). *Proceedings of the 1st International Conference on Education Data Mining.* Montreal, 2008.

[66] Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI journal)*. 19(3), 2009.

[67] Feng, M, Heffernan, N., Heffernan, C. & Mani, M. (2009). Using mixed-effects modeling to analyze different grain-sized skill models. *IEEE Transactions on Learning Technologies*, vol. 2, no. 2, pp. 79-92.

[68] Feng, M., Heffernan, N., & Beck, J. (2009). Using learning decomposition to analyze instructional effectiveness in the ASSISTment system. In Dimitrova, Mizoguchi, du Boulay, & Graesser (Eds), *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-2009).* Amsterdam, Netherlands: IOS Press. Brighton, UK.

[69] Feng, M., Beck, J., Heffernan, N.T. (2009). Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning. In Barnes & Desmarais (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining*. Cordoba, Spain, 2009.

[70] Feng, M., Beck, J. (2009). Back to the future: a non-automated method of constructing transfer models. In Barnes & Desmarais (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining*. Cordoba, Spain, 2009.

[71] Ferguson, K., Arroyo, I., Mahadevan,S., Woolf, B., and Barto, A. (2006). Improving Intelligent Tutoring Systems: Using Expectation Maximization to Learn Student Skill Levels. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Berlin: Springer-Verlag. pp.453-462.

[72] Ferguson, K., Woolf, B., & Mahadevan,S. (2009). Transfer learning and representation discovery in intelligent tutoring systems. In Dimitrova, Mizoguchi, du Boulay, & Graesser (Eds), *Artificial Intelligence in Education*. Amsterdam, Netherlands: IOS Press. pp. 605-607.

[73] Fitzmaurice, G., Laird, N., & Ware, J. (2004). Applied Longitudinal Analysis. Hoboken, New Jersey: Wiley & Sons.

[74] Gierl, M.J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in Algebra on the SAT. *Journal of Technology, Learning, and Assessment*, 6(6). Retrieved May, 2008 from http://www.jtla.org.

[75] Grigorenko, E. L. and Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124,* 75–111.

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

[77] Heffernan, N. T. & Koedinger, K.R. (1997). The composition effect in symbolizing: The role of symbol production vs. text comprehension. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum. pp. 307-312.

[78] Heffernan, N. T. & Koedinger, K. R. (1998). A developmental model for algebra symbolization: The results of a difficulty factors assessment. In M. Gernsbacher & S. Derry (Eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum. pp. 484-489.

[79] Heffernan N.T., Turner T.E., Lourenco A.L.N., Macasek M.A., Nuzzo-Jones G., Koedinger K.R.(2006). The ASSISTment Builder: Towards an Analysis of Cost Effectiveness of ITS creation. FLAIRS2006, Florida, USA (2006).

[80] Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, 6(3), 249-260.

[81] Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika*, 55(3): 357:373.

[82] Junker, B.W. & Sijtsma K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

[83] Junker, B. (2007). Using on-line tutoring records to predict end-of-year exam scores: experience with the ASSISTments project and MCAS 8th grade mathematics. In Lissitz, R. W. (Ed.), *Assessing and modeling cognitive development in school*. Maple Grove, MN: JAM Press.

[84] Kay, J. (1997). Learner Know Thyself: Student Models to Give Learner Control and Responsibility, in Z. Halim, T. Ottomann & Z. Razak (eds),

Proceedings of International Conference on Computers in Education, Association for the Advancement of Computing in Education (AACE), 17-24.

[85] Katz, S., Lesgold, A., Eggan, G., & Gordin, M. (1992). Modeling the Student in Sherlock II, *The International Journal of Artificial Intelligence in Education*, 3(4), 495-518.

[86] Klein Entink, R.H., Fox, J.-P., & van der Linden, W. J. (2008). A multivariate multilevel approach to simultaneous modeling of accuracy and speed on test items. *Psychometrika, 73*.

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

[88] Koedinger, K. (2000). Research statement for Dr. Kenneth R. Koedinger, June, 2000. Retrieved from http://pact.cs.cmu.edu/koedinger/koedingerReserach.html.

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

[90] Koedinger, K. R., Anderson, J. R., Hadley, W. H., &Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30–43.

[91] Koedinger, K. R., Aleven, V., Heffernan. N. T., McLaren, B. & Hockenberry, M. (2004). Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, pages 162-173, Maceio, Brazil.

[92] Lee, C. & Ting, D. (2005). Predicting MCAS score from the Assistment system. Project report. Department of Statistics, Carnegie Mellon University. Available at http://www.cs.wpi.edu/~mfeng/pub/TingLeePaper.pdf

[93] Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41, 205-237.

[94] Leighton, J. & Gierl, M. (Eds.) (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. New York, NY: Cambridge University Press.

[95] Leszczenski, J. M., & Beck, J. E. (2007, July 9). What's in a word? Extending learning factors analysis to modeling reading transfer. *Proceedings*

of the AIED2007 Workshop on Educational Data Mining, Marina del Rey, CA, 31-39.

[96] Massachusetts Department of Education. (2000). Massachusetts Mathematics Curriculum Framework. Retrieved from http://www.doe.mass.edu/frameworks/math/2000/final.pdf, Nov. 6th, 2005.

[97] MCAS technical report (2001). Retrieved from http://www.cs.wpi.edu/mfeng/pub/mcas_techrpt01.pdf, August 5th, 2005.

[98] McCalla, G. I. and Greer, J. E. (1994). Granularity- based reasoning and belief revision in student models. In Greer, J. E. and McCalla, G. I., (eds), *Student Modeling: The Key to Individualized Knowledge-Based Instruction*, pages 39-62. Springer-Verlag, Berlin.

[99] Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). Implementing Data-Informed Decision Making in Schools--Teacher Access, Supports and Use. Report prepared for U.S. Department of Education, Office of Planning, Evaluation and Policy Development. Prepared by SRI International, Menlo Park, CA. Jan, 2009.

[100] Militello, M., Sireci, S., & Schweid, J. (2008). Intent, purpose, and fit: An examination of formative assessment systems in school districts. Paper presented at the American Educational Research Association, New York City, NY.

[101] Mislevy, R.J. (1995). Probability-based reasoning in cognitive diagnosis. Chapter 3 in P. Nichols, S. Chipman, & R. Brennan, eds., *Cognitively Diagnostic Assessment,* Hillsdale, NJ: Erlbaum.

[102] Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

[103] Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

[104] Mislevy, R. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.) *Educational measurement* (4th ed.). Washington, DC: American Council on Education.

[105] Mitchell, T. (1997). *Machine Learning.* Columbus, OH: McGraw-Hill.

[106] Mitrovic, A. & Martin, B. (2002). Evaluating the Effects of Open Student Models on Learning. In P. De Bra, P. Brusilovsky & R. Conejo (eds), *Adaptive Hypermedia and Adaptive Web-Based Systems, Proceedings of*

Second International Conference, Springer-Verlag, Berlin Heidelberg, 296-305.

[107] Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In P. Feltovich (Ed.), *Smart Machines in Education*. Menlo Park, CA: MIT/AAAI Press. pp. 169-234.

[108] Murray, R.C., VanLehn, K. & Mostow, J. (2001). A decision-theoretic approach for selecting tutorial discourse actions. In E. Horvitz, T. Paek, & C. Thompson (Eds.), *Proceedings of the NAACL Workshop on Adaptation in Dialogue Systems*, Pittsburgh, PA, June, 2001, pp. 41-48. New Brunswick, NJ: Association for Computational Linguistics.

[109] Nathan, M. J., and Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction*, 18(2), 209-237.

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

[111] Nichols, P. D., Chipman, S. F., & Brennan, R.L. (1995) (Eds.). *Cognitively Diagnostic Assessment*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[112] No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425. (2002). Retrieved September 6, 2005, from http://www.ed.gov/legislation/ESEA02/

[113] Nuzzo-Jones, G., Walonoski, J.A., Heffernan, N.T. & Livak, T. (2005). The eXtensible tutor architecture: A new foundation for ITS. Workshop on Adaptive Systems for Web-Based Education: Tools and Reusability held at the 12th Annual Conference on Artificial Intelligence in Education. Amsterdam. pp. 1-7.

[114] Olson, L. (2004). State Test Programs Mushroom as NCLB Mandate Kicks In *Education Week,* Nov. 20th, pp. 10-14.

[115] Olson, L. (2005). Special report: testing takes off. *Education Week,* November 30, 2005, pp. 10–14. Also available on-line from http://www.edweek.org/media/13testing.pdf

[116] Pardos, Z. A., Feng, M., Heffernan, N. T. & Heffernan, C. L. (2007). Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effect Methods. *Proceedings of the 13th Conference on Artificial Intelligence In Education* (AIED 2007).

[117] Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. L. (2006). *Using Fine Grained Skill Models to Fit Student Performance with Bayesian Networks.* Workshop in Educational Data Mining held at the Eighth International Conference on Intelligent Tutoring Systems. Taiwan. 2006.

[118] Pardos, Z. & Heffernan, N. (2009a). Detecting the Learning Value of Items in a Randomized Problem Set. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-2009).* Amsterdam, Netherlands: IOS Press. Brighton, UK.

[119] Pardos, Z. & Heffernan, N.T. (2009b). Determining the Significance of Item Order In Randomized Problem Sets. In *Proc. of the 2nd International Conference on Educational Data Mining*. Cordoba, Spain.

[120] Patvarczki, J., Almeida, S., Beck,J. & Heffernan, N.T. (2008) Lessons Learned from Scaling Up a Web-Based Intelligent Tutoring System. In Woolf & Aimeur (Eds.) *Proceeding of the 9th International Conference on Intelligent Tutoring Systems. Springer-Verlag: Berlin.*

[121] Patvarczki, J., Politz J., Heffernan, N. (2009). Scalability and Robustness in the Domain of Web Based Tutoring. Scalability issues in AIED Workshop at the 14th International Conference on Artificial Intelligence in Education.

[122] Pavlik, P.I., Cen, H. & Koedinger, K.R. (2009a). Performance factors analysis – A new alternative to knowledge tracing. In Dimitrova, Mizoguchi, du Boulay, & Graesser (Eds), *Artificial Intelligence in Education*. Amsterdam, Netherlands: IOS Press. pp.531-538.

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

[124] R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

[125] Raftery, A. E. (1995). Bayesian model selection in social research. In *Sociological Methodology*, 25, 111-163.

[126] Ramírez, E., & Clark, K. (Feb., 2009). What Arne Duncan Thinks of No Child Left Behind: The new education secretary talks about the controversial law and financial aid forms. (Electronic version). Retrieved on March 8th, 2009 from http://www.usnews.com/articles/education/2009/02/05/what-arne-duncan-thinks-of-no-child-left-behind.html.

[127] Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D.Wright. Chicago: The University of Chicago Press.

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

[129] Razzaq, L., Heffernan, N.T. (2006). Scaffolding vs. hints in the Assistment System. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 635-644. 2006.

[130] Razzaq, L., Heffernan, N.T., Lindeman, R.W. (2007). What level of tutor interaction is best?. In Luckin & Koedinger (Eds.). *Proceedings of the 13th Conference on Artificial Intelligence in Education*. Amsterdam, Netherlands: IOS Press.

[131] Razzaq, L., Heffernan, N., Feng, M., Pardos, Z. (2007). Developing Fine-Grained Transfer Models in the ASSISTment System. *Journal of Technology, Instruction, Cognition, and Learning* , Vol. 5. Number 3. Old City Publishing, Philadelphia, PA. 2007. pp. 289-304.

[132] Razzaq, L., Parvarczki, J., Almeida, S.F., Vartak, M., Feng, M., Heffernan, N.T. and Koedinger, K. (2009). The ASSISTment builder: Supporting the Life-cycle of ITS Content Creation. *IEEE Transactions on Learning Technologies*, 07 May. 2009. IEEE computer Society Digital Library.

[133] Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. Cognitive Science, 21, 1-29.

[134] Ritter, S. and Koedinger, K. (1996). An architecture for plug-in tutor agents. Journal of Artificial Intelligence in Education, 7(3-4):315-347, 1996.

[135] Ritter, S., Harris, T., Nixon, T., Dickison, D., Murray, C., Towle, B. (2009). Reducing the knowledge tracing space. In Barnes, Desmarais, Romero, & Ventura (Eds.). *Proceedings of the 2nd International Conference on Educational Data Mining*. pp. 151-160. Cordoba, Spain.

[136] Roediger, H.L. III, and Karpicke, J.D. (2006). The power of testing memory. *Perspectives on Psychological Science*. 1(3), pp. 181- 210.

[137] Rothman, S. (2001). 2001 MCAS Reporting Workshop: The second generation of MCAS results. Massachusetts Department of Education. Downloaded November 2006 from http://www.doe.mass.edu/mcas/2001/news/reporting wkshp.pps.

[138] Russell, M., & Masters, J. (2009). Formative Assessment Tools for Algebra and Geometry. Presentation at Technology Supports for Formative Assessment Symposium of 2009 American Educational Research Association (AERA) Annual Meeting. San Diego, CA.

[139] Sao Pedro, M., Gobert, J., Heffernan, N. & Beck, J. (2009). Comparing Pedagogical Approaches for Teaching the Control of Variables Strategy. Cognitive Science Society Annual 2009 Conference.

[140] Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and occurrence*. Oxford University Press, New York.

[141] Sternburg, R.J., & Grigorenko, E.L. (2001). All testing is dynamic testing. *Issues in Education*, 7, 137-170.

[142] Sternburg, R.J., & Grigorenko, E.L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, England: Cambridge University Press.

[143] Stiggins, R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324-328.

[144] Tan, E. S., Imbos, T. & Does R. J. M. (1994) A distribution-free approach to comparing growth of knowledge. *Journal of Education Measurement,* 31 (1):51-65.

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

[146] Tatsuoka, K.K. (1990). Toward an Integration of Item Response Theory and Cognitive Error Diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

[147] Turner, T. E., Macasek, M. A., Nuzzo-Jones, G., Heffernan, N. T. (2005). The ASSISTment Builder: A Rapid Development Tool for ITS. In *Proceedings of the 12th Annual Conference on Artificial Intelligence in Education*, pages 929-931.

[148] United States Department of Education. (2003). Using data to influence classroom decisions. Washington, DC.

[149] van der Linden, W. J. & Hambleton, R. K. (Eds.) (1997). Handbook of Modern Item Response Theory. New York: Springer-Verlag.

[150] van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*. 287-308.

[151] van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5-20.

[152] VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence and Education*, 15(3), 1-47.

[153] Walonoski, J., Heffernan, N.T. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 382-391. 2006.

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

[155] Wiliam, D. (2006). Formative Assessment: Getting the focus right. *Educational Assessment,* 11(3&4), 283-289.

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

[157] Wongchokprasitti, C., & Brusilovsky, P. (2007). NewsMe: A Case Study for Adaptive News Systems with Open User Model. In: Proceedings of The Third International Conference on Autonomic and Autonomous Systems, ICAS 2007, Athens, Greece, June 19-25, 2007, IEEE Press.

[158] Wylie, E. C., & Ciofalo, J. (2008). Supporting teachers' use of individual diagnostic items. Teachers College Record. Retrieved from http://www.tcrecord.org/PrintContent.asp?ContentID=15363 on October 13, 2008.

[159] Yun, J.T., Willet. J. & Murnane, R. (2004) Accountability-Based Reforms and Instruction: Testing Curricular Alignment for Instruction Using the Massachusetts Comprehensive Assessment System. Paper presented at the Annual American Educational Research Association Meeting. San Diego.

[160] Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2005). BILOG-MG 3 – Multiple-Group IRT Analysis and Test maintenance for Binary Items.

Scientific Software International, Inc., Lincolnwood, IL. URL http://www.ssicentral.com/.

Appendices

Appendix A. List of Publications

Book chapters

- **B2. Feng, M.,** Heffernan, N.T., & Koedinger, K.R. (in Submission). Student modeling in an Intelligent Tutoring System. Submitted to Stankov, Glavinc, and Rosic. (Eds.) Intelligent Tutoring Systems in E-learning Environments: Design, Implementation and Evaluation. IGI Global. early 2010 (anticipated)
- **B1.** Razzaq, L., **Feng, M.**, Heffernan, N., Koedinger, K., Nuzzo-Jones, G., Junker, B., Macasek, M., Rasmussen, K., Turner, T., & Walonoski, J. (2007). Blending Assessment and Instructional Assistance. In Nedjah, Mourelle, Borges and Almeida (Eds). *Intelligent Educational Machines within the Intelligent Systems Engineering Book Series* . pp.23-49. Springer Berlin / Heidelberg. (pdf)

Journal papers

- **J6.** Razzaq, L., Parvarczki, J., Almeida, S.F., Vartak, M., **Feng, M.**, Heffernan, N.T. and Koedinger, K. (2009). The ASSISTment builder: Supporting the Life-cycle of ITS Content Creation. *IEEE Transactions on Learning Technologies*. 2(2) 157-166.
- **J5. Feng, M**, Heffernan, N., Heffernan, C. & Mani, M. (2009). Using mixed-effects modeling to analyze different grain-sized skill models. *IEEE Transactions on Learning Technologies*, 2(2), pp. 79-92. (pdf) (Based on **WP3**) (**Featured article of the issue**)

- **J4. Feng, M.**, Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an Online System that tutors as it assesses. In *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*. 19(3), 243-266, August, 2009. (pdf) (Based on **CP2**)
- **J3.** Razzaq, L., Heffernan, N., **Feng, M.**, Pardos, Z. (2007). Developing Fine-Grained Transfer Models in the ASSISTment System. *Journal of Technology, Instruction, Cognition, and Learning* , Vol. 5. Number 3. Old City Publishing, Philadelphia, PA. 2007. pp. 289-304.(pdf) (Based on **WP3**)
- **J2. Feng, M.** & Heffernan, N. (2007). Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in the Assistment System. *Journal of Interactive Learning Research.* 18 (2), pp. 207-230. Chesapeake, VA: AACE. (pdf) (Based on **J1, WP1**)
- **J1. Feng, M.**, Heffernan, N.T. (2006). Informing Teachers Live about Student Learning: Reporting in the Assistment System. *Technology, Instruction, Cognition, and Learning* Journal. Vol. 3. Old City Publishing, Philadelphia, PA. 2006. (pdf [preview version]) (Based on **WP1**)

Conference papers

- **CP9. Feng, M**., Beck, J., Heffernan, N.T. (2009). Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning. In Barnes, Desmarais, Romero, & Ventura (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining*. pp. 51-60. Cordoba, Spain: Copisterias Don Folio, S.L. (pdf)
- **CP8. Feng, M**., Heffernan, N.T., Beck, J. (2009). Using learning decomposition to analyze instructional effectiveness in the ASSISTment system. In Dimitrova, Mizoguchi, du Boulay, and Grasser (Eds), *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-2009)*. pp. 523-530. Amsterdam, Netherlands: IOS Press. (pdf)
- **CP7.** Hansen E. G., Zapata-Rivera, D., & **Feng, M.** (2009). Beyond Accessibility: Evidence Centered Design for Learning (ECDL) for Improving the Efficiency of Instruction. Paper presented at the session of Test use in special populations at the National Council on Educational Measurement 2009 Annual Conference (NCME, 2009), San Diego, CA. pdf
- **CP6. Feng, M.,** Heffernan, N., Beck, J, & Koedinger, K. (2008) Can we predict which groups of questions students will learn from? In Baker & Beck (Eds.). *Proceedings of the 1st International Conference on Education Data Mining*. pp.218-225. Montreal 2008.(pdf)

- **CP5. Feng, M.**, Beck, J,. Heffernan, N. & Koedinger, K. (2008) Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test? In Baker & Beck (Eds.). *Proceedings of the 1st International Conference on Education Data Mining*. pp.107-116. Montreal 2008. (pdf) (Based on **CP2**)
- **CP4. Feng, M.**, Heffernan, N. T. (2007). Assessing Students. Performance Longitudinally: Item Difficulty Parameter vs. Skill Learning Tracking. Paper presented at the 2007 Annual meeting of National Council of Measurement on Educational (NCME'2007), Chicago. (pdf) (Based on **WP3**)
- **CP3. Feng, M.**, Heffernan, N.T, Koedinger, K.R. (2006b). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 31-40. 2006. (pdf)
- **CP2. Feng, M.**, Heffernan, N.T, Koedinger, K.R. (2006a) Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. In *Proceedings of the 15th International World Wide Web Conference*. pp. 307-316. New York, NY: ACM Press. 2006. (pdf) **Best Student Paper Nominee**.
- **CP1.** Razzaq, L., **Feng, M.**, Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, pp. 555-562. Amsterdam: ISO Press. (pdf)

Poster papers in prestigious conferences

- **PP2. Feng, M**., Beck, J. (2009). Back to the future: a non-automated method of constructing transfer models. In Barnes & Desmarais (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining*. pp. 240-249. Cordoba, Spain: Copisterias Don Folio, S.L. (pdf)
- **PP1.** Pardos, Z., **Feng, M.** & Heffernan, N. T. & Heffernan-Lindquist, C. (2007).Analyzing fine-grained skill models using bayesian and mixed effect methods. In Luckin & Koedinger (Eds.) Proceedings of the 13th Conference on Artificial Intelligence in Education. Amsterdam, Netherlands: IOS Press.pp.626-628. (pdf) (Based on **WP3, WP4**)

Workshop and less stringently reviewed venues

- **WP5. Feng, M.**, Hansen, E. & Zapata, D. (2009). Using Evidence Centered Design for Learning (ECDL) to examine ASSISTment. Paper presented at the annual meeting of America Educational Research Association (AERA), San Diego, CA. April, 2009. pdf
- **WP4**. Pardos, Z., **Feng, M.**, Heffernan, N. T., Heffernan-Lindquist, C.& Ruiz, C. (2007). Analyzing fine-grained skill models using bayesian and mixed effect methods. In Heiner, C., Heffernan, N., & Barnes, T. (Eds). *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*. Marina del Rey, CA. pp. 50-59.
- **WP3. Feng, M.**, Heffernan, N.T, Mani, M. & Heffernan C. (2006). Using Mixed-Effects Modeling to Compare Different Grain-Sized Skill Models. In Beck, J., Aimeur, E., & Barnes, T. (Eds). *Educational Data Mining: Papers from the AAAI Workshop*. Menlo Park, CA: AAAI Press. pp. 57-66. (pdf)
- **WP2. Feng, M.**, Heffernan, N.T., Koedinger, K.R., (2005). Looking for Sources of Error in Predicting Student's Knowledge. In Beck. J. (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Menlo Park, California: AAAI Press. pp. 54-61. (pdf)
- **WP1. Feng, M.**, Heffernan, N.T. (2005). Informing Teachers Live about Student Learning: Reporting in the Assistment System. The 12th International Conference on Artificial Intelligence in Education Workshop on Usage Analysis in Learning Systems, 2005, Amsterdam.(pdf)

Appendix B.

Sample MCAS test items from year 2003

Grade 8 Mathematics Test

The spring 2003 Grade 8 MCAS Mathematics Test was based on learning standards in the Massachusetts *Mathematics Curriculum Framework (2000)*. The *Framework* identifies the five major content strands listed below.

- Number Sense and Operations
- Patterns, Relations, and Algebra
- Geometry
- Measurement
- Data Analysis, Statistics, and Probability

The grade 7–8 learning standards for each of these strands appear on pages 62–66 of the *Mathematics Curriculum Framework*, which is available on the Department website at www.doe.mass.edu/frameworks/math/2000/final.pdf.

In *Test Item Analysis Reports* and on the *Subject Area Subscore* pages of the MCAS *School Reports* and *District Reports*, Mathematics test results are reported under five MCAS reporting categories, which are identical to the five *Mathematics Curriculum Framework* content strands listed above.

Test Sessions and Content Overview

The grade 8 Mathematics Test contained two separate test sessions. Each session included multiple-choice and open-response questions. Session 1 also included short-answer questions. Common test items are shown on the following pages as they appeared in test booklets.

Reference Materials and Tools

During testing, each student taking the Grade 8 Test was provided with a *Grade 8 Mathematics Reference Sheet* and a plastic ruler. A copy of this reference sheet follows the final question in this chapter.

While answering questions during Session 2, each student had sole access to a calculator with at least four functions and a square root key. Calculator use was not allowed during Session 1. No other reference tools or materials were allowed, with the exception of bilingual word-to-word dictionaries used by limited English proficient students.

Cross-Reference Information

The table at the conclusion of this chapter indicates each item's reporting category and the *Framework* learning standard it assesses. The correct answers for multiple-choice and short-answer questions are also displayed in the table.

Mathematics

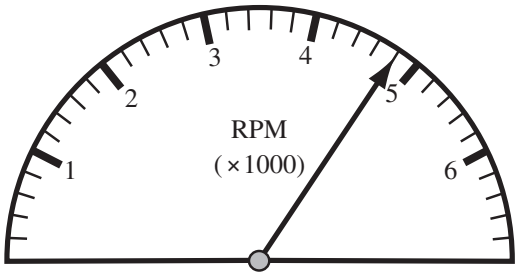
SESSION 1

You may use your reference sheet during this session.
*You may **not** use a calculator during this session.*

DIRECTIONS

This session contains fifteen multiple-choice questions, five short-answer questions, and two open-response questions. Mark your answers to these questions in the spaces provided in your Student Answer Booklet.

1 The figure below shows an RPM gauge.



How many RPMs is the gauge registering?

A. 44

B. 48

C. 4080

D. 4800

2 Which of the following statements is **false**?

A. $(-12)(-12)(-12) = 3(-12)$

B. $12(4 - 1) = 12(4) - 12(1)$

C. $12 + (4 - 1) = (12 + 4) - 1$

D. $-12 + 12 = 12 + (-12)$

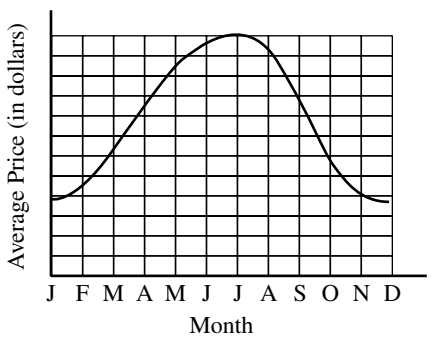
3 The chart below shows the average monthly price per share of HiTek stock for each month in 2001.

2001 HiTek Stock Prices

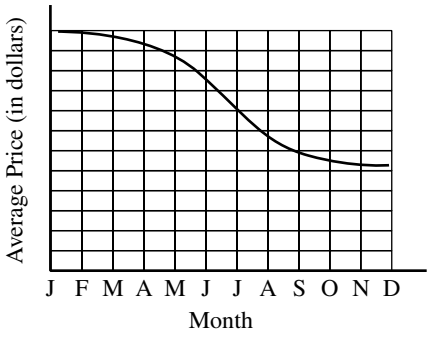
Month	January	February	March	April	May	June	July	August	September	October	November	December
Average Price (in $)	22.61	24.25	31.02	27.31	29.92	33.10	36.14	35.50	34.01	31.05	36.20	40.12

Which of the following curves **best** models the general behavior of the stock's price for last year?

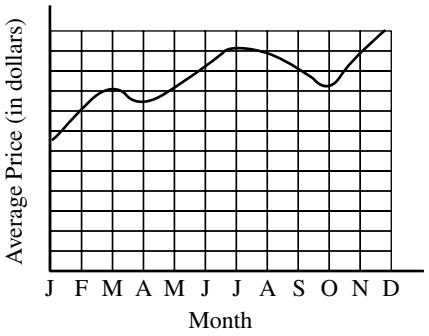
A.



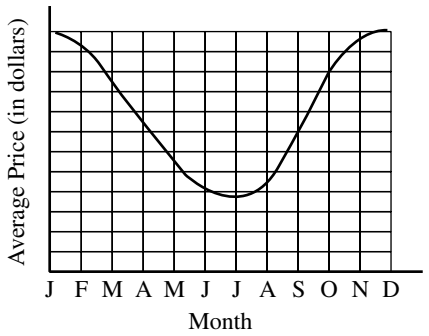
C.



B.



D.



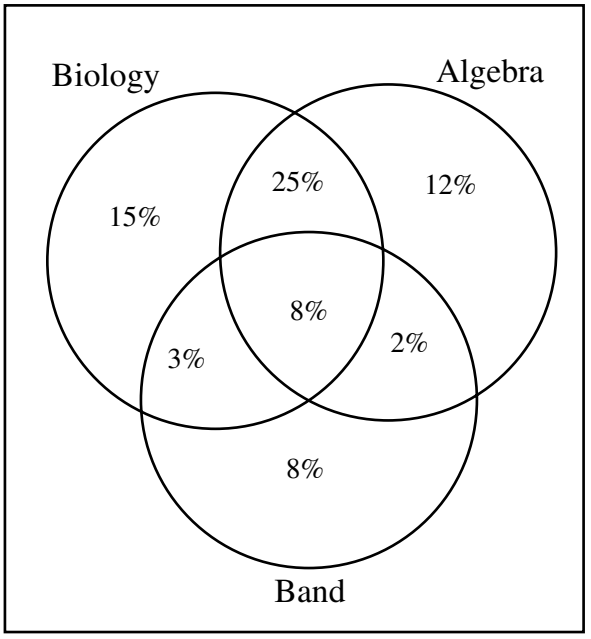
Question 9 is an open-response question.

- **BE SURE TO ANSWER AND LABEL ALL PARTS OF THE QUESTION.**
- **Show all your work (diagrams, tables, or computations) in your Student Answer Booklet.**
- **If you do the work in your head, explain in writing how you did the work.**

Write your answer to question 9 in the space provided in your Student Answer Booklet.

9 At student registration, eighth-grade students selected the courses they would be taking next year as ninth graders. The counselor made the diagram below that shows a relationship among the percentages of students who chose to take Biology, Algebra, and/or Band.

Student Registration



a. According to the diagram, what percent of the eighth-grade students will be taking **all three** courses, Biology, Algebra, and Band, next year?

b. What percent of the eighth-grade students will be taking Algebra and Biology, but **not** Band, next year?

c. If 900 students signed up to take courses, how many will **not** be taking Biology, Algebra or Band? Show or explain your work.

Questions 19, 20, and 21 are short-answer questions. Write your answers to these questions in the boxes provided in your Student Answer Booklet. Do not write your answers in this test booklet. You may do your figuring in the test booklet.

19 Triangles *ABC* and *DEF* shown below are congruent.

The perimeter of $\triangle ABC$ is 23 inches. What is the length of side $\overline{DF}$ in $\triangle DEF$?

20 What is $\frac{3}{4}$ of $1\frac{1}{2}$?

21 Write a rule that could be used to show the relationship between x and y in the table below.

x	y
-4	16
-1	1
0	0
3	9
7	49

Massachusetts Comprehensive Assessment System
Grade 8 Mathematics Reference Sheet

Use the information below and your ruler as needed to answer questions in this test.

PERIMETER FORMULAS

square.....$P = 4s$

rectangle .....$P = 2b + 2h$

triangle.....$P = a + b + c$

CIRCLE FORMULAS

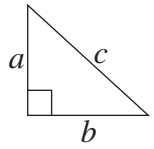
circle .....$C = 2\pi r$

OR

$C = \pi d$

$A = \pi r^2$

PYTHAGOREAN THEOREM



$a^2 + b^2 = c^2$

CONVERSIONS

1 mile = 5280 feet

1 square mile = 640 acres

AREA FORMULAS

square.....$A = s^2$

rectangle.....$A = bh$

OR

$A = lw$

triangle .....$A = \frac{1}{2}bh$

circle .....$A = \pi r^2$

trapezoid .....$A = \frac{1}{2}h(b_1 + b_2)$

VOLUME FORMULAS

rectangular prism .....$V = Bh$
(B = area of base)

cone.....$V = \frac{1}{3}\pi r^2 h$

cylinder .....$V = \pi r^2 h$

cube.....$V = s^3$
(s = length of an edge)

Grade 8 Mathematics
Spring 2003 Released Items:
Reporting Categories, Standards, and Correct Answers

Item No.	Page No.	Reporting Category	Standard	Correct Answer (MC/SA)*
1	170	*Measurement*	8.M.1	D
2	170	*Number Sense and Operations*	8.N.8	A
3	171	*Data Analysis, Statistics, and Probability*	8.D.2	B
4	172	*Patterns, Relations, and Algebra*	8.P.7	C
5	172	*Data Analysis, Statistics, and Probability*	8.D.1	A
6	172	*Data Analysis, Statistics, and Probability*	8.D.3	C
7	173	*Patterns, Relations, and Algebra*	8.P.6	–2
8	173	*Data Analysis, Statistics, and Probability*	8.D.2	20 years old
9	174	*Data Analysis, Statistics, and Probability*	8.D.2	
10	175	*Patterns, Relations, and Algebra*	8.P.2	B
11	175	*Patterns, Relations, and Algebra*	8.P.7	A
12	175	*Patterns, Relations, and Algebra*	8.P.5	B
13	176	*Patterns, Relations, and Algebra*	8.P.7	D
14	176	*Number Sense and Operations*	8.N.4	C
15	176	*Data Analysis, Statistics, and Probability*	8.D.3	D
16	177	*Geometry*	8.G.7	B
17	177	*Patterns, Relations, and Algebra*	8.P.7	D
18	177	*Patterns, Relations, and Algebra*	8.P.5	C
19	178	*Geometry*	8.G.2	10 inches
20	178	*Number Sense and Operations*	8.N.12	$\frac{9}{8}$ or equivalent
21	178	*Patterns, Relations, and Algebra*	8.P.1	$y = x^2$ or $x^2 = y$
22	179	*Geometry*	8.G.3	
23	180	*Number Sense and Operations*	8.N.9	B
24	180	*Number Sense and Operations*	8.N.10	B
25	180	*Data Analysis, Statistics, and Probability*	8.D.3	C
26	180	*Number Sense and Operations*	8.N.3	B
27	181	*Patterns, Relations, and Algebra*	8.P.1	D
28	182	*Patterns, Relations, and Algebra*	8.P.1	
29	183	*Number Sense and Operations*	8.N.10	
30	184	*Number Sense and Operations*	8.N.2	B
31	184	*Geometry*	8.G.1	C
32	184	*Data Analysis, Statistics, and Probability*	8.D.4	C
33	184	*Number Sense and Operations*	8.N.9	A
34	185	*Measurement*	8.M.2	C
35	185	*Number Sense and Operations*	8.N.3	D
36	185	*Patterns, Relations, and Algebra*	8.P.1	A
37	186	*Patterns, Relations, and Algebra*	8.P.9	B
38	186	*Measurement*	8.M.3	C
39	187	*Measurement*	8.M.3	

* Answers are provided here for multiple-choice and short-answer items only. Sample responses and scoring guidelines for open-response items, which are indicated by shaded cells, will be posted to the Department's website later this year.

Appendix C.

MCAS Curriculum Framework for 8th grade mathematics

Learning Standards by Grade Span or Course for Grades 7–12

NUMBER SENSE AND OPERATIONS

Understand numbers, ways of representing numbers, relationships among numbers, and number systems

Understand meanings of operations and how they relate to one another

Compute fluently and make reasonable estimates

GRADES 7–8 LEARNING STANDARDS

Students engage in problem solving, communicating, reasoning, connecting, and representing as they:

8.N.1 Compare, order, estimate, and translate among integers, fractions and mixed numbers (i.e., rational numbers), decimals, and percents.

8.N.2 Define, compare, order, and apply frequently used irrational numbers, such as $\sqrt{2}$ and π.

8.N.3 Use ratios and proportions in the solution of problems, in particular, problems involving unit rates, scale factors, and rate of change.

8.N.4 Represent numbers in scientific notation, and use them in calculations and problem situations.

8.N.5 Apply number theory concepts, including prime factorization and relatively prime numbers, to the solution of problems.

8.N.6 Demonstrate an understanding of absolute value, e.g., $|-3| = |3| = 3$.

8.N.7 Apply the rules of powers and roots to the solution of problems. Extend the Order of Operations to include positive integer exponents and square roots.

8.N.8 Demonstrate an understanding of the properties of arithmetic operations on rational numbers. Use the associative, commutative, and distributive properties; properties of the identity and inverse elements (e.g., $-7 + 7 = 0$; $\frac{3}{4} \times \frac{4}{3} = 1$); and the notion of closure of a subset of the rational numbers under an operation (e.g., the set of odd integers is closed under multiplication but not under addition).

8.N.9 Use the inverse relationships of addition and subtraction, multiplication and division, and squaring and finding square roots to simplify computations and solve problems, e.g. multiplying by $\frac{1}{2}$ or 0.5 is the same as dividing by 2.

8.N.10 Estimate and compute with fractions (including simplification of fractions), integers, decimals, and percents (including those greater than 100 and less than 1).

8.N.11 Determine when an estimate rather than an exact answer is appropriate and apply in problem situations.

8.N.12 Select and use appropriate operations—addition, subtraction, multiplication, division, and positive integer exponents—to solve problems with rational numbers (including negatives).

Understand patterns, relations, and functions

Represent and analyze mathematical situations and structures using algebraic symbols

Use mathematical models to represent and understand quantitative relationships

Analyze change in various contexts

GRADES 7–8 LEARNING STANDARDS

Students engage in problem solving, communicating, reasoning, connecting, and representing as they:

8.P.1 Extend, represent, analyze, and generalize a variety of patterns with tables, graphs, words, and, when possible, symbolic expressions. Include arithmetic and geometric progressions, e.g., compounding.

8.P.2 Evaluate simple algebraic expressions for given variable values, e.g., $3a^2 - b$ for $a = 3$ and $b = 7$.

8.P.3 Demonstrate an understanding of the identity $(-x)(-y) = xy$. Use this identity to simplify algebraic expressions, e.g., $(-2)(-x+2) = 2x - 4$.

8.P.4 Create and use symbolic expressions and relate them to verbal, tabular, and graphical representations.

8.P.5 Identify the slope of a line as a measure of its steepness and as a constant rate of change from its table of values, equation, or graph. Apply the concept of slope to the solution of problems.

8.P.6 Identify the roles of variables within an equation, e.g., $y = mx + b$, expressing y as a function of x with parameters m and b.

8.P.7 Set up and solve linear equations and inequalities with one or two variables, using algebraic methods, models, and/or graphs.

8.P.8 Explain and analyze—both quantitatively and qualitatively, using pictures, graphs, charts, or equations—how a change in one variable results in a change in another variable in functional relationships, e.g., $C = \pi d$, $A = \pi r^2$ (A as a function of r), $A_{rectangle} = lw$ ($A_{rectangle}$ as a function of l and w).

8.P.9 Use linear equations to model and analyze problems involving proportional relationships. Use technology as appropriate.

8.P.10 Use tables and graphs to represent and compare linear growth patterns. In particular, compare rates of change and x- and y-intercepts of different linear patterns.

GEOMETRY

Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships

Specify locations and describe spatial relationships using coordinate geometry and other representational systems

Apply transformations and use symmetry to analyze mathematical situations

Use visualization, spatial reasoning, and geometric modeling to solve problems

GRADES 7–8 LEARNING STANDARDS

Students engage in problem solving, communicating, reasoning, connecting, and representing as they:

8.G.1 Analyze, apply, and explain the relationship between the number of sides and the sums of the interior and exterior angle measures of polygons.

8.G.2 Classify figures in terms of congruence and similarity, and apply these relationships to the solution of problems.

8.G.3 Demonstrate an understanding of the relationships of angles formed by intersecting lines, including parallel lines cut by a transversal.

8.G.4 Demonstrate an understanding of the Pythagorean theorem. Apply the theorem to the solution of problems.

8.G.5 Use a straight-edge, compass, or other tools to formulate and test conjectures, and to draw geometric figures.

8.G.6 Predict the results of transformations on unmarked or coordinate planes and draw the transformed figure, e.g., predict how tessellations transform under translations, reflections, and rotations.

8.G.7 Identify three-dimensional figures (e.g., prisms, pyramids) by their physical appearance, distinguishing attributes, and spatial relationships such as parallel faces.

8.G.8 Recognize and draw two-dimensional representations of three-dimensional objects, e.g., nets, projections, and perspective drawings.

MEASUREMENT

Understand measurable attributes of objects and the units, systems, and processes of measurement

Apply appropriate techniques, tools, and formulas to determine measurements

GRADES 7–8 LEARNING STANDARDS

Students engage in problem solving, communicating, reasoning, connecting, and representing as they:

8.M.1 Select, convert (within the same system of measurement), and use appropriate units of measurement or scale.

8.M.2 Given the formulas, convert from one system of measurement to another. Use technology as appropriate.

8.M.3 Demonstrate an understanding of the concepts and apply formulas and procedures for determining measures, including those of area and perimeter/circumference of parallelograms, trapezoids, and circles. Given the formulas, determine the surface area and volume of rectangular prisms, cylinders, and spheres. Use technology as appropriate.

8.M.4 Use ratio and proportion (including scale factors) in the solution of problems, including problems involving similar plane figures and indirect measurement.

8.M.5 Use models, graphs, and formulas to solve simple problems involving rates, e.g., velocity and density.

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them

Select and use appropriate statistical methods to analyze data

Develop and evaluate inferences and predictions that are based on data

Understand and apply basic concepts of probability

GRADES 7–8 LEARNING STANDARDS

Students engage in problem solving, communicating, reasoning, connecting, and representing as they:

8.D.1 Describe the characteristics and limitations of a data sample. Identify different ways of selecting a sample, e.g., convenience sampling, responses to a survey, random sampling.

8.D.2 Select, create, interpret, and utilize various tabular and graphical representations of data, e.g., circle graphs, Venn diagrams, scatterplots, stem-and-leaf plots, box-and-whisker plots, histograms, tables, and charts. Differentiate between continuous and discrete data and ways to represent them.

8.D.3 Find, describe, and interpret appropriate measures of central tendency (mean, median, and mode) and spread (range) that represent a set of data. Use these notions to compare different sets of data.

8.D.4 Use tree diagrams, tables, organized lists, basic combinatorics ("fundamental counting principle"), and area models to compute probabilities for simple compound events, e.g., multiple coin tosses or rolls of dice.