Intelligent Tutoring Systems in E-Learning Environments:

Design, Implementation and Evaluation

Slavomir Stankov
University of Split, Croatia

Vlado Glavinić
University of Zagreb, Croatia

Marko Rosić
University of Split, Croatia

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Director of Book Publications:	Julia Mosemann
Acquisitions Editor:	Lindsay Johnston
Development Editor:	Julia Mosemann
Publishing Assistant:	Travis Gundrum
Typesetter:	Deanna Jo Zombro
Production Editor:	Jamie Snavely
Cover Design:	Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Intelligent tutoring systems in e-learning environments : design, implementation, and evaluation / Slavomir Stankov, Vlado Glavinic and Marko Rosic, editors.
p. cm.
Includes bibliographical references and index.
Summary: "This book addresses intelligent tutoring system (ITS) environments from the standpoint of information and communication technology (ICT) and the recent accomplishments within both the e-learning paradigm and e-learning systems"--Provided by publisher.   ISBN 978-1-61692-008-1 (hardcover) -- ISBN 978-1-61692-009-8 (ebook) 1. Intelligent tutoring systems. 2. Web-based instruction--Design. 3. Education--Effect of technological innovations on. I. Stankov, Slavomir, 1947- II. Glavinic, Vlado, 1952- III. Rosic, Marko, 1970-   LB1028.73.I65 2010
371.33'44678--dc22
2009052667

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 11
Student Modeling in an Intelligent Tutoring System

Mingyu Feng
SRI International, USA

Neil Heffernan
Worcester Polytechnic Institute, USA

Kenneth Koedinger
Carnegie Mellon University, USA

ABSTRACT

Student modeling and cognitively diagnostic assessment are important issues that need to be addressed for the development and successful application of intelligent tutoring systems (its). Its needs the construction of complex models to represent the skills that students are using and their knowledge states, and practitioners want cognitively diagnostic information at a finer grained level. This chapter reviews our effort on modeling student's knowledge in the ASSISTment project. Intelligent tutors have been mainly used to teach students. In the ASSISTment project, we have emphasized using the intelligent tutoring system as an assessment system that provides instructional assistance during the test. Usually it is believed that assessment get harder if students are allowed to learn during the test, as its then like try to hit a moving target. So our results are surprising that by providing tutoring to students while they are assessed we actually prove the assessment of students' knowledge. Additionally, in this article, we present encouraging results about a fine-grained skill model with that system that is able to predict state test scores. We conclude that using intelligent tutoring systems to do assessment seems like a reasonable way of dealing with the dilemma that every minute spent testing students takes time away from instruction.

INTRODUCTION

In the United States there are concerns about poor student performance on new high-stakes standards based tests required by the No Child Left Behind Act (NCLB, 2002) legislation. To address this issue, educational technologies, like intelligent tutoring systems (ITS) have been developed and proven to be useful helping students learn. For instance, the Cognitive Tutors leads to large learning gains (Koedinger et al, 1997). Recently, President Obama made a commitment

DOI: 10.4018/978-1-61692-008-1.ch011

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, …, *learning software as effective as a personal tutor*…." [1] So, how do we build ITS that are as effective as a personal tutor? To create an effective piece of learning software, we need a good model of student learning. 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 build 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.

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 said 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's progress in real-time, so teachers and parents can assess and monitor student strengths and weaknesses. To reflect on the "continuous assessment" idea, we 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, thus, obtaining a better understanding of each student's pre-existing, or current knowledge status and how that knowledge is changing over time.

Student models in ITS are constructive, especially cognitive models. 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 cognitively diagnostic assessment and teacher reporting. Wiliam (2006) describes an assessment as formative only if information about what is being assessed results in change that would otherwise not occur. Here the definition of formative assessment becomes more detailed to the utility for teachers to understand if and how students are learning. By giving teachers cognitively diagnostic data in a timely fashion, teachers can change their teaching in response to the data they collect about student understanding. The US 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.

In this chapter, we will describe how various student modeling approaches have been applied in an online cognitively diagnostic assessment system, called the ASSISTment System that provides both assistance and assessment in an integrated fashion. In the second section, we focus on giving an overview of the ASSISTment System, including the structure of an ASSISTment, the problem sequencing, the teacher reports, the authoring tools, content development and usage and also the evidence showing the effectiveness of tutoring in ASSISTments. The third section of this chapter is devoted to student modeling in ITS. We first conduct a literature review of student modeling approaches, and then report our work in the ASSISTment System. We will describe how we improve the accuracy of assessment by measuring how much assistance students need and by tracking the change of student knowledge longitudinally. Furthermore, our fine-grained cognitive models that map each question to a few knowledge components allow us to more accurately predict these scores. We conclude the chapter with some general implications in the forth section.

BACKGROUND ON THE ASSISTMENT SYSTEM

There has been intense interest in using periodic benchmark tests to predict student performance on end-of-year accountability assessments (Olson, 2005). Some teachers make extensive use data from practice tests and released items from statewide assessments to target specific student knowledge needs and identify learning opportunities for individual students as well as the class as a whole so that their instruction could be data-driven. However, such formative assessments not only require great effort and dedication, but also take valuable time away from instruction. Limited classroom time in middle school classes compels teachers to choose between time spent assisting students' learning and time spent assessing students' abilities. Critics of NCLB 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. But does it have to be? What if we better integrated assessment into the classroom and we allowed students to learn during the test? Noticing the situation, Dr. Heffernan at Worcester Polytechnic Institute and his colleagues at Carnegie Mellon University started to build a system ("ASSISTments"[2], http://www.assistment.org) to help resolve this dilemma.

Traditionally the 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. Therefore, in order to put them together the solution here must involve a way whereby students can take an assessment, but at the same time, make sure that students' time is spent primarily on learning. It should be able to provide an accurate prediction of student performance on the standardized tests so that teachers have an idea of how their students will perform on end-of-year assessment. Meanwhile, it will present a more fine-grained evaluation of student abilities so the teachers will be able to use this detailed feedback to tailor their instruction to focus on the particular difficulties identified by the system. The ASSISTment project's[3] goal is to provide cognitive-based assessment of students while providing tutoring content to students. The system aims to assist students in learning the different skills needed for the Massachusetts Comprehensive Assessment System (MCAS[4]) test or (other state tests) while at the same time assessing student knowledge to provide teachers with fine-grained assessment of their students' knowledge. The system tutors students in learning different skills using scaffolding questions, hints, and incorrect messages (or buggy messages) (Razzaq et al., 2005). Assessment of student performance is provided to teachers through real-time reports based on statistical analysis.

The system is primarily used by middle and high school[5] teachers and students throughout Massachusetts who are preparing for the MCAS test. Using the web-based ASSISTment System is free and only requires registration on the website; no software need be installed. The system started in the year 2004 by interviewing a few middle school math teachers about how they would tutor a problem. Then when the system was built and ready to be used, these teachers were willing to bring in their students to try it before the MCAS. Those were the first 60 students who worked in ASSISTments. Because the ASSISTment System was practical, easily-accessible, informative, and more importantly free, it was gradually accepted by more and more teachers and schools.

Things changed from Heffernan and colleagues running around schools to introduce the system to school administrators and teachers contact us to "join" the ASSISTment team. Currently, there are over 3000 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 and 17 teachers are participating in a workshop at WPI to learn how to make more effective use of diagnostic assessment data from ASSISTments. Heffernan and his colleagues including WPI graduate students still visit schools regularly to help with technical issues and on how to use ASSISTment data to drive classroom instructions. It is worth pointing out that both the ASSISTment System and its authoring tools are open, generic and not limited to any specific domain. The system was initially built to help 8th grade (13 to 14 year olds) students be better prepared for the MCAS. The subject content was mainly mathematics, as this was the subject on which students from Worcester, Massachusetts showed worst performance[6]. Later on, more content such as 6th grade and 10 grade mathematics materials were introduced into the system. More recently, a new program was launched in the ASSISTment project to tutor science inquiry skills (Sao Pedro, Gobert, Heffernan & Beck, 2009).

Though the ASSISTment System is a derivative of the Cognitive Tutor (Anderson et al., 1995), it is built for different classroom use than the Cognitive Tutor. Cognitive Tutor students are intended to use the tutor for two class periods a week. Students are expected to proceed at their own rate letting the mastery learning algorithm advance them through the curriculum. Some students will make steady progress while others will be stuck on early units. There is value in this in that it allows students to proceed at their own paces. One downside from the teachers' perspective could be that they might want to have their class all do the same material on the same day so they can assess their students simultaneously. The ASSISTment System was created with this classroom use in mind. The ASSISTment System was created with the idea that teachers would use it once every two weeks as part of their normal classroom instruction, meant more as a formative assessment system and less as the primary means of assessing students. Cognitive Tutor advances students only after they have mastered all of the skills in a unit. We know that some teachers use some features to automatically advance students to later lessons because they might want to make sure all the students get some practice on quadratics, for instance.

We think that no one system is "the answer" but that they have different strengths and weaknesses. If the student uses the computer less often there comes a point where mastery learning based program (i.e., the Cognitive Tutor) may be behind on what a student knows, and seem to move along too slowly to teachers and students. On the other hand, a weakness of ASSISTments is that does not offer mastery learning and adaptive activity selection, so if students struggle, it does not automatically adjust. It is assumed that the teacher (and not the computer system) will decide if a student needs to go back and look at a topic again.

The Structure of an ASSISTment

Koedinger et al. (2004) introduced pseudo-tutors which mimic cognitive tutors. The ASSISTment system uses a further simplified pseudo-tutor, called an ASSISTment, where only a linear progression through a problem is supported which makes content creation easier and more accessible to a general audience.

An ASSISTment consists of a single main question (a.k.a. original question) and a tutoring session for assistance. As students work in the system, the main question will be presented first and can be treated as an assessment task for which students need to submit an answer. In contrast to a traditional testing environment, students can request assistance if they don't know how to answer the question, though it is generally thought to be pedagogically more desirable that a student submits a thoughtful answer before accessing the tutoring. For any given problem, assistance to students is available either in the form of a hint sequence or a set of scaffolding questions. Hints are messages that provide insights and suggestions for solving a specific problem, and each hint sequence ends with a *bottom-out* hint that gives the student the answer. Scaffolding questions are designed to lead the student one-step-at-a-time to the solution and each step addresses specific skills needed to answer the original question. Students must answer each scaffolding question in order to proceed to the next scaffolding question. When students finish all of the scaffolding questions, they may be presented with the original question again to finish the problem. Each scaffolding question also has a hint sequence to help the students answer the question if they need extra help. Additionally, constructive feedback called *buggy* messages are provided to students if certain anticipated incorrect answers are selected or entered, otherwise a generic feedback message will be shown. For problems without scaffolding, a student will remain in a problem until the problem is answered correctly and can ask for hints that are presented one at a time. If scaffolding is available, the student will be programmatically advanced to the scaffolding questions in the event of an incorrect answer. The ASSISTments assume that students may know certain skills and rather than slowing them down by going through all of the scaffolding questions first, ASSISTment allowS students to try to answer questions without showing every step. This differs from Cognitive Tutors (Anderson et al., 1995) and Andes (VanLehn et al., 2005) which both ask the students to fill in many different steps in a typical problem.

Underlying Architecture

The system was built on the eXtensible Tutor Architecture (XTA) (Nuzzo-Jones et al., 2005), a framework that controls the interface and behaviors of our intelligent tutoring system via a collection of modular units. Each conceptual unit has an abstract and extensible implementation allowing for evolving tutor types and content delivery methods. Different from many other ITSs that are client-based such as the Cognitive Tutors and Andes, the ASSISTment System, is a web-based tutoring system that allows us to bypass the process of client software installation and administration, thus, drastically increases the accessibility and visibility of the system, as suggested by Ritter & Koedinger (1996) and Brusilovsky & Peylo (2003). It also provides great control over content distribution, server updates, and log data collection. Meanwhile, the server-centered structure also brings up scalability issue. Patvarczki et al. (2008) and Patvarczki, J., Politz J., Heffernan, N. (2009) addressed the scalability issue by introducing a load-balancing and fault- tolerance architecture. They proposed a symmetric clustered infrastructure to avoid single-points-of-failure, and, increase the fault- tolerance of ASSISTments. More over, a load-balancing algorithm was developed to distribute the load between the cluster members based on round-robin selection. These works build foundation for further expansion of the ASSISTment System for a broader adoption.

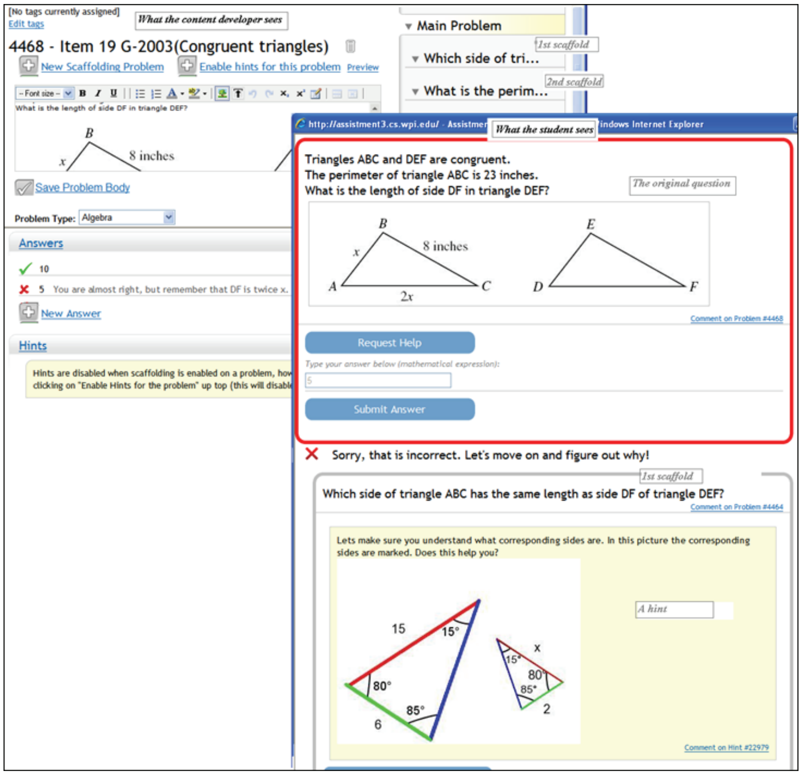
Authoring Tools

Hints, scaffolds, and buggy messages together help create ASSISTments that are structurally simple but can address complex student behavior and provide appropriate intervention. The structure and the supporting interface used to build ASSISTments (*the authoring tools* or sometimes referred to as *the builder,*Razzaq et al., 2009), shown in Figure 1, uses common web technologies such as

HTML and JavaScript, allowing it to be used on most modern browsers. The authoring tools are simple enough that users with little or no computer programming experience or cognitive psychology background can use it easily. Figure 1 shows an ASSISTment being built on the left and what the student sees is shown on the right. Content authors can easily enter question text, hints and buggy messages by clicking on the appropriate field and typing; formatting tools are also provided for easily bolding, italicizing, etc. Images and animations can also be uploaded in any of these fields. The builder also enables scaffolding within scaffold questions, although this feature has not been used often in our existing content.

Several studies (Heffernan et al., 2006; Turner et al., 2005) have been conducted to evaluate the authoring tools in terms of usability and decreased creation time of tutors. The builder was augmented to track how long it takes authors to create an ASSISTment[7]. Thus, once we know how many ASSISTments authors have created, we can estimate the amount of content tutoring time created by using the previously established number that students spend about 2 minutes per ASSISTment (Heffernan et al., 2006). This produced a ratio of development time to on-line instruction time of about 40:1, comparing against the literature suggesting a 200:1 ratio (Anderson et al., 1995). Therefore, our method for creating tutoring content was much more cost effectively. We did this by building a tool that reduces both the skills needed to create content as well as the time needed to do so. Our subject matter expert was satisfied with the quality of the contents. Additionally, as we will point out later in section 2.1.6, several studies showed that the ASSISTments created by the content authors produced significant learning.

Figure 1. The builder and the corresponding student screen



[No tags currently assigned]

4468 - Item 19 G-2003(Congruent triangles)

What is the length of side DF in triangle DEF?

Problem Type: Algebra

✓ 10

✗ 5 You are almost right, but remember that DF is twice x.

Hints are disabled when scaffolding is enabled on a problem, how
clicking on "Enable Hints for the problem" up top (this will disabl

Triangles ABC and DEF are congruent.
The perimeter of triangle ABC is 23 inches.
What is the length of side DF in triangle DEF?

Type your answer below (mathematical expression):

5

✗ Sorry, that is incorrect. Let's move on and figure out why!

Which side of triangle ABC has the same length as side DF of triangle DEF?

Lets make sure you understand what corresponding sides are. In this picture the corresponding sides are marked. Does this help you?

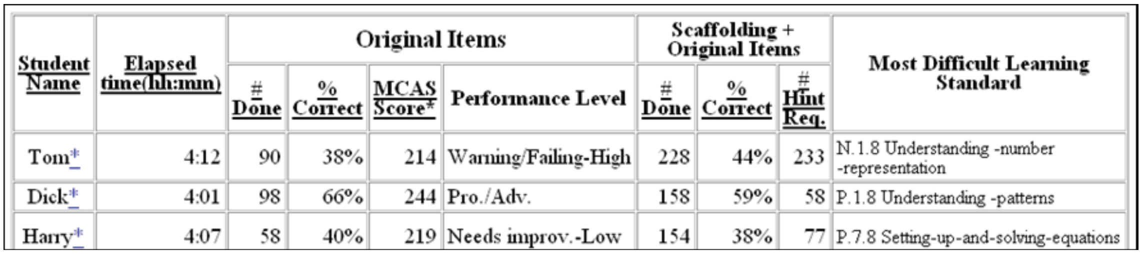
Reporting

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. But teachers do not want to wait six months for the state to grade the exams. Teachers and parents also want better feedback than they currently receive. The reporting (Feng & Heffernan, 2007b) in the ASSISTment System has been 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 this detailed feedback to tailor their instruction to focus on the particular difficulties identified by the system.

The "Grade Book," shown in Figure 2, is the ASSISTment report used most frequently by teachers. Each row in the report represents information for one student, including our prediction of his MCAS score based on student response to the original questions. Besides presenting information on the item level, it also summarizes the student's actions in "ASSISTment metrics" that tells more about students' actions besides their performance. For example, it illuminates students' unusual behaviour, such as making far more attempts and requesting more hints than other students in the class. By clicking the link of the skill that the student has the lowest percent correct, 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. Finding out which knowledge components are difficult for students may also help us improve our item sequencing strategies.

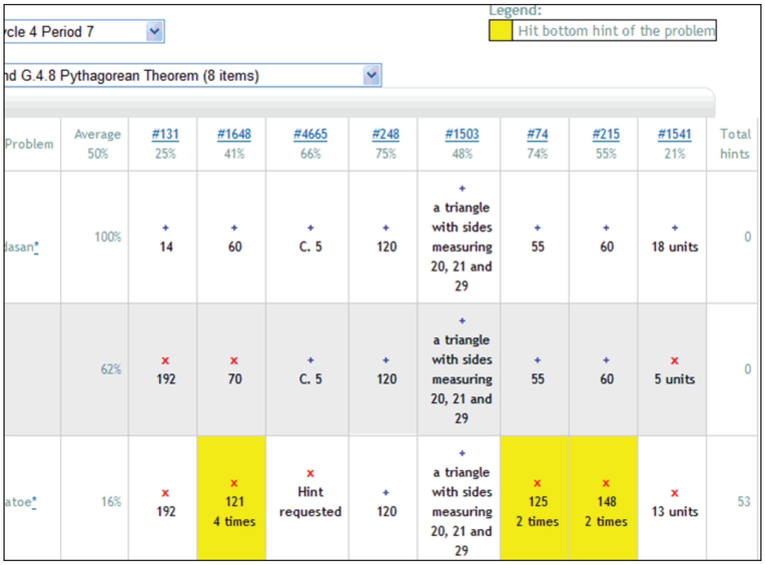
The grade book report gives an overview of a student/a class's performance. Figure 3 shows an item report that is generated based on detailed, action-level logs of students. The report shows teachers how students are doing on individual problems. In particular, the item report in Figure 3 shows how three students from one class did on an assignment of eight problems focusing on Pythagorean Theorem, a skill that students need to acquire in the eighth grade. By presenting information 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), or incorrectly (indicated by an "x" sign) at their first attempt, or requested a hint (indicated by a message "Hint requested"). For instance, the three yellow-highlighted cells in the third row leap out at the first sight. Therefore, 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 provide better remediation to students.

Figure 2. Grade book report



Student Name	Elapsed time(hh:mm)	Original Items				Scaffolding + Original Items			Most Difficult Learning Standard
		# Done	% Correct	MCAS Score*	Performance Level	# Done	% Correct	# Hint Req.	
Tom*	4:12	90	38%	214	Warning/Failing-High	228	44%	233	N.1.8 Understanding -number -representation
Dick*	4:01	98	66%	244	Pro./Adv.	158	59%	58	P.1.8 Understanding -patterns
Harry*	4:07	58	40%	219	Needs improv.-Low	154	38%	77	P.7.8 Setting-up-and-solving-equations

Figure 3. An item report tells teachers how students are doing on individual problems



ʏcle 4 Period 7

Legend:
🟨 Hit bottom hint of the problem

ıd G.4.8 Pythagorean Theorem (8 items)

Problem	Average 50%	#131 25%	#1648 41%	#4665 66%	#248 75%	#1503 48%	#74 74%	#215 55%	#1541 21%	Total hints
ıasan*	100%	+ 14	+ 60	+ C. 5	+ 120	+ a triangle with sides measuring 20, 21 and 29	+ 55	+ 60	+ 18 units	0
	62%	x 192	x 70	+ C. 5	+ 120	+ a triangle with sides measuring 20, 21 and 29	+ 55	+ 60	x 5 units	0
atoe*	16%	x 192	x 121 4 times	x Hint requested	+ 120	+ a triangle with sides measuring 20, 21 and 29	x 125 2 times	x 148 2 times	x 13 units	53

Teachers 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.

Content Development and Management

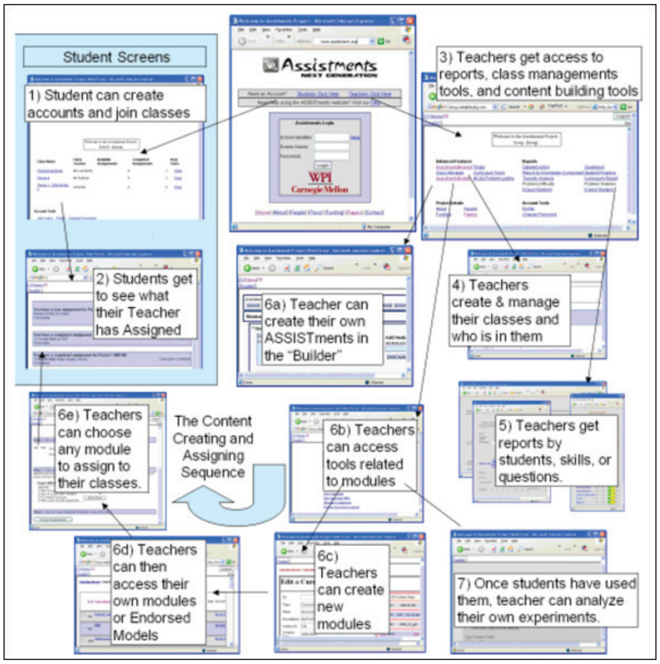
We are attempting to support the full life cycle of content authoring and management with the tools available in the ASSISTment system. Teachers can create problems with tutoring, map each question to the skills required to solve them, bundle problems together in sequences that students work on, view reports on students' work and use tools to maintain and refine their content over time.

Figure 4 shows how 1) students login, 2) get assignments to do that show up such as in the right hand side of Figure 1. Figure 2 also shows that our web-based system allows teachers access to 3) get reports, 4) manage classes, 5) get reports on students, 6a) create, edit and maintain content with the builder, 6b) find their own and others people's content (such as their students' content) 6c-e) bundling that content and assigning it to their students. We even have working reports (step 7) that automatically analyze the results of experiments that randomly assign students to conditions, which is the sort of analysis we need to determine if learning is happening.

Analyzing Learning Effectiveness in ASSISTments

We analyzed data within the ASSISTment System usage to determine whether the System effectively teaches. For the studies reported by Feng, Heffernan, & Koedinger (2006a, 2006b), we used the ASSISTment System to track student knowledge longitudinally over the course of a schools year, based upon each student using our system about a dozen times during the course of the year. This result confounded learning from the computer system with students learning from their sitting their normal class. To eliminate this confound, Feng, Heffernan, Beck & Koedinger (2008) looked to see if students were reliably learning from their time spent with the *computer in a single day.* We conducted a focused analysis of a subset

Figure 4. ASSISTment supports the full life cycle of content authoring


Student Screens
1) Student can create accounts and join classes
Assistments
3) Teachers get access to reports, class managements tools, and content building tools
2) Students get to see what their Teacher has Assigned
6a) Teacher can create their own ASSISTments in the "Builder"
4) Teachers create & manage their classes and who is in them
6e) Teachers can choose any module to assign to their classes.
The Content Creating and Assigning Sequence
6b) Teachers can access tools related to modules
5) Teachers get reports by students, skills, or questions.
6d) Teachers can then access their own modules or Endorsed Models
6c) Teachers can create new modules
7) Once students have used them, teacher can analyze their own experiments.

of items. 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). We assessed learning 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 tended to perform better on later opportunities of items in a GLOP, then they might have learned[8] from the instructional assistance provided on items by the ASSISTment system that they worked on earlier by answering the scaffolding questions or by reading hint messages. Our results suggested that students performed better later in the same computer session on similar skills, which indicated students learned from using ASSISTments. However, learning was rather uneven across groups of skills. We brought up a few hypotheses to explain this phenomenon and found out that students learned more from group of items that are more cohesive. Interestingly, human expert judgments were more predictive as to which groups of skills were learnable than the automated approaches. Following this work, Feng, Beck, & Heffernan (2009a, 2009b) and Pardos & Heffernan (2009a) reported our efforts on analyzing the relative effectiveness of items in a GLOP. More recently, Pardos et al. (2009) engaged in finding prerequisite relationship among problems in a GLOP using educational data mining approach. The results that have been validated through simulation studies provided an opportunity to improve overall learning from a GLOP and to inform content authors of item pairs of high transfer.

As seen from our student survey results, students complained that being forced into scaffolding questions was time consuming and frustrating. We were not sure if all of the time we invested into these "fancy" scaffolding questions was worth it. Luckily, the ASSISTment System provides a good platform to run randomized controlled experiments to find answers for such questions. Razzaq & Heffernan (2006) and Razzaq, Heffernan, & Lindeman (2007) reported the experiments comparing different tutoring strategies, hints vs. scaffolding. vs. delayed feedback[9]. The results showed that in general, scaffolding led to higher averages on a post-test, although it was not statistically significant. After a closer examination of the effect of math proficiency and the level of feedback on learning, they found out that honor students benefit more from delayed feedback and regular students did better in the scaffolding condition. One possible explanation was that less proficient students benefit from more interaction and coaching through each step to solve a problem while more proficient students benefit from seeing problems worked out and seeing the big picture. The fact that one strategy seems better than the other is an important research finding that potentially could improve the overall instructional efficacy of the system. However, that is not the main point. The point is that it is a future goal for the ASSISTment System to do this sort of analysis automatically for content creators, and based on the research results, pick the most suitable tutoring for the individual learners.

STUDENT MODELING

Literature Review of Student Modeling Approaches

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. Modeling student response data from intelligent tutoring systems has a long history (e.g. Corbett, Anderson, & O'Brien, 1995; Draney, Pirolli, & Wilson, 1995). 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. In addition, they found complex skills can be decomposed into subskills that 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).

Corbett and his colleagues employed a very detailed model of skills, but their system did not have questions tagged with more than one production rule (Anderson, 1993). Our collaborators (Ayers & Junker, 2006) were engaged in trying to allow multi-mapping[10] using a version of the fine grained model but reported their Linear Logistic Test Model (LLTM) does not fit well. Different from the approach we will report later in this chapter, the model they applied does not track student performance over time.

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 feedback to students and teachers based on evidence. Almond et al. (2007) examined the application of Bayesian frameworks to Item Response Theory-based cognitive diagnostic modeling. Bayesian networks have also been used to investigate the results of skill hierarchies using real world data in intelligent tutoring systems (e.g., Ferguson et al., 2006) and simulated users (e.g., Collins, Greer, & Huang, 1996; Daniel et al., 2007). 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 to provide tailored support (VanLehn, 2005). 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, for example, how to make inferences about a group of students. To ease the complicated coding effort involved in the usage of Bayes nets, especially 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 between student knowledge and observed behavior. Later on in Beck et al. (2008) dynamic Bayesian networks were 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, & 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 & 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 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 work integrated the educational benefits of allowing students to access the learner model content, 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 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).

There has also been a large interest in building cognitively diagnostic models. What we refer to as a "skill model" later in section 3.2 is referred to as "Q-Matrix" by some Artificial Intelligence researchers (Barnes, 2005) and psychometricians (Tatsuoka, 1990); Croteau, Heffernan, & Koedinger (2004) called it "transfer model"; while Cen, Koedinger & Junker (2005), and Griel, 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 have been using automatic/semi-automatic techniques to search for skill models, including the rule space method (Tatsuoka, 1990), the Q-matrix method (Barnes, 2005), and Learning Factor Analysis (Cen, Koedinger, & Junker, 2005, 2006).

In the following section, we will describe our effort on modeling students in the context of ASSISTment System.

Student Modeling in ASSISTments

The ASSISTment System has two assessment goals: predicting student performance on end-of-year accountability exams, and cognitively assess student knowledge to help teachers target next instructional steps. These goals are complicated by two features of the system: assessment is ongoing throughout the school year as student proficiency develops, and the ASSISTment System itself as a tutoring system changes student proficiency. Nevertheless, prediction using simple student models and a number of "assistance metrics" (summaries of hint-seeking behavior, time spent on questions correctly vs. incorrectly answered, etc.) can be almost optimally effective at predicting end-of-year exam scores, as we will show below. Statistical uncertainty in teacher feedback reports based on more-detailed student models is sometimes surprisingly low, but even in cases where the per-student uncertainty is high, reports aggregated over groups of students can be quite reliable. These ideas will be considered both for the ASSISTment System and in the broader context of online assessment and learning systems.

Predicting Student End-of-Year Exam Score

The first assessment goal for ASSISTments is to make a prediction of student end-of-year exam scores. We have reported the results of several studies where we used log data collected via ASSISTments to try to predict scores of the MCAS high-stake tests required of all students in the state of Massachusetts. We will now review of a few of those results. All the participants of studies mentioned in this chapter were from Worcester Public Schools, Worcester, Massachusetts. They were all 8th grade students at the time they used ASSISTments. The students practiced problems within the ASSISTment System as a part of their regular math class. No selection processes based on student demographic information or personal traits have been conducted before the data analysis. In all of these studies we compared students actual MCAS test scores with the predicted to get calculate the Mean Absolute Deviation (MAD) which was then used as the measure to evaluate the student models.

As mentioned above, the ASSISTment System is not a pure assessment system but also provides instructional assistance when students have difficulties. Does providing assistance hurt the accuracy of the assessment? Surprisingly, these studies reported that the assistance provided actually improves the assessment. The idea is that by seeing how much help students needed allows a more sensitive measure of student knowledge then just whether they got a question correct.

Much work has been done in the past 10 years or so on developing "online testing metrics" for dynamic testing (or dynamic assessment) (Grigorenko & Sternberg, 1998) to supplement accuracy data (wrong/right scores) for characterizing student proficiency. Researchers have been interested in trying to get more assessment value by comparing traditional assessment (students getting an item marked wrong or even getting partial credit) with a measure that shows how much help they needed. Brown, Bryant, & Campione (1983) compared traditional testing paradigms against a dynamic testing paradigm. Grigorenko & 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. In this study they wanted to predict learning gains between pretest and posttest. They found that static testing prediction did not correlate (R = 0.45) with student learning data as well as the "dynamic testing" did (R = 0.60). Brown et al. (1983) suggested that this method could be effectively done by computer, but, as far as we know, their work was not continued. Luckily, the ASSISTment system provides an ideal test bed as it already provides a set of hints to students. So it is a natural way to extend and test this idea and see if we can replicate their finding of ASSISTment-style measures being better assessors. Our hypothesis was that we could 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 learn how to solve a test item.

We continued the dynamic testing approach (Feng, Heffernan & Koedinger, 2006a, 2006b, 2009; Feng, Beck, Heffernan, & Koedinger, 2008) and developed a group of "assistance" metrics that focused on student-system-interaction behaviors during the tutoring session (scaffoldings or hints). The metrics included accuracy (percent correct on scaffolding questions), speed (how many seconds a student needs to solve a problem), attempts (on average, how many attempts a student made to finally get the correct answer) and help-seeking behavior (on average how often a student asks for hint messages; how many times he/she reaches the bottom-out hint). These metrics were either not available or have largely been ignored in traditional assessment environments. But as a computer-based tutoring system, ASSISTments have the potential to use far more. We computed the metrics from our log data and found out that all these metrics were reliably correlated with students' real MCAS test scores. Then we built different students models to predict end-of-year MCAS scores. Our goal was to see if we can reliably predict students' test scores and to evaluate how well on-line use of the ASSISTment system can help in the prediction.

A number of different regression models were compared for measuring student knowledge during ASSISTments use, including a static model that was based only on accuracy on original questions, a dynamic assessment model that was built on all assistance metrics except the accuracy, and a mixed model that incorporates the static model and the dynamic model. The key contrast of interest was between a static model that mimicked paper practice tests by scoring students as either correct or incorrect on each main question, with the dynamic assessment model that leveraged the assistance metrics to take into account the amount of assistance students need before they get an item correct. We trained the models on two data sets from two school years. The data set we collected during the school year of 2004-2005 involved 417 students who practiced mathematics problems in the ASSISTment system since September 2004 till May 2005, for a mean length of 267 minutes (standard deviation = 79) across about 9 class sessions, finishing on average 147 items (standard deviation = 60). For this data set, the predicted score from the static model correlated with the MCAS test scores, with an R-value of 0.733 and the dynamic assistance model correlated with an R-value of 0.821, reliably higher than the correlation between the static prediction and MCAS scores. The second data set came from the usage of 616 8th grade students during 2005-2006 school year. These 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). Again, we found the same trend in the data set that the dynamic assessment model did reliably better on predicting MCAS scores (R = 0.816) than the static model (R = 0.784). For both data sets, we could improve our prediction of MCAS score further by combining the assistance metrics with the static model. Additionally, we verified the validity of the combined model by cross-validating the combined model both inside same year's data and using data across two years. Thus, we can claim that the ASSISTment System did a better job of predicting student knowledge by being able to take into consideration metrics such as how much tutoring assistance was needed and how fast a student solves a problem.

We suspect that a better job of predicting MCAS scores could be done if students could be encouraged to take the system seriously and reduce "gaming behavior." One way to reduce gaming is to detect it and then to notify the teacher (in the teacher's reporting session) with evidence that the teacher can use to approach the student. Our preliminary work on gaming detection was presented in Walonoski & Heffernan (2006). It is assumed that teacher intervention will lead to reduced gaming behavior, and thereby affect more accurate assessment, and higher learning. Adding visual feedback, as one ongoing work in the ASSISTment System does, aims to help teachers quickly detect gaming behaviors.

Tracking Student Performance Longitudinally

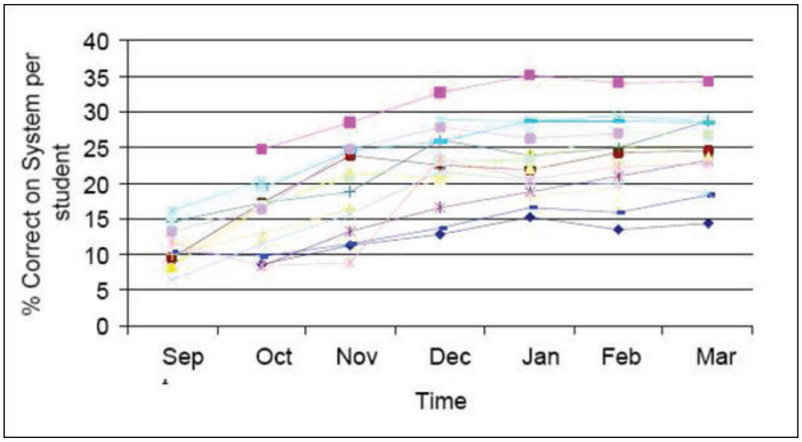
In Razzaq et al. (2005) and Feng, Heffernan, Beck, & Koedinger (2008), we reported results that suggested students were learning directly during the assisting in ASSISTments. We did this by looking at groups of items that had the same skills and to see if performance later in the class period was associated with higher 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, we focused on tracking student progress that results from both classroom instruction and ASSISTments tutoring over a long periods of time. To investigate this question, we did a 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. We gradually introduced factors such as what school they are in, who their teacher is, or which class they are from, into our models. By doing so, we attempted to provide an answer to the question of what factors impact (or are correlated with) students' learning rate.

During the school year of 2004-2005, there were approximately 650 students using the system, with each student coming to the computer lab about 7 times. We created a table with 4550 rows, one row for each student for each day, with an average percent correct which itself is averaged over about 15 MCAS items done on a given day. In Figure 5, average student performance is plotted versus time. The y-axis is the average percent correct on the original item (student performance on the scaffolding questions is ignored in this analysis) in a given class. The x-axis represents time, where data is combined together into months, so some students who came to the lab twice in a month will have their numbers averaged. The fact that most of the class trajectories are generally rising suggests that most classes are learning between months. The result of statistical modeling confirms what we saw in the plot. Our fitted longitudinal model ended up with a statistically significant learning slope that indicates that student performance was reliably increasing during the school year. Additionally, our model was able to detect different rates of learning at different schools, but not among different teachers and classes.

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, or students were not really learning but were being tested on easier items. In the future, this confound will be eliminated by sampling items randomly. Adding automated applied longitudinal data analysis is being pursued.

More work is needed to build models to better detect differences between teachers' effects on the learning rates of students that presumably exist. Besides this, other factors will be investigated about their possible impact on students' learning over time. Information from student profiles such as gender, race and ethnicity, special education status, free-lunch status, etc., is being added. During this analysis, we noticed the fact that

Figure 5. Average student performance is plotted over time



generally speaking, groups with higher estimated initial scores showed lower rates of learning. Our 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 10th grade and won't appear in the ASSISTment system. Further investigation needs to be done to explain this phenomenon.

Modeling Fine Grained Student Knowledge

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 are not good at *informing educators* about which skills are causing difficulty and how to help students. However, the question of tagging items with learning standards is very important because schools 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. In this section, we describe our work on constructing fine-grained skill models. More precisely, our research question is what the right grain size of skill models and how model granularity impacts the effectiveness of a student's knowledge tracking. Tracking student skills at a very fine- grained level is not applicable in the traditional assessment environment because of limited testing time and test items where 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. Original questions are always followed by scaffolding questions, each addressing a single piece of the student's knowledge. Therefore, when a student answered the original question wrong, we can rely on his responses to scaffolding questions to figure out exactly where the student has a misunderstanding and are able to track the specific knowledge component precisely. Moreover, students are using the ASSISTment 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. The continuous usage allows us to collect more evidence of student's performance on every skill.

Skill Models Development and Skill Mapping

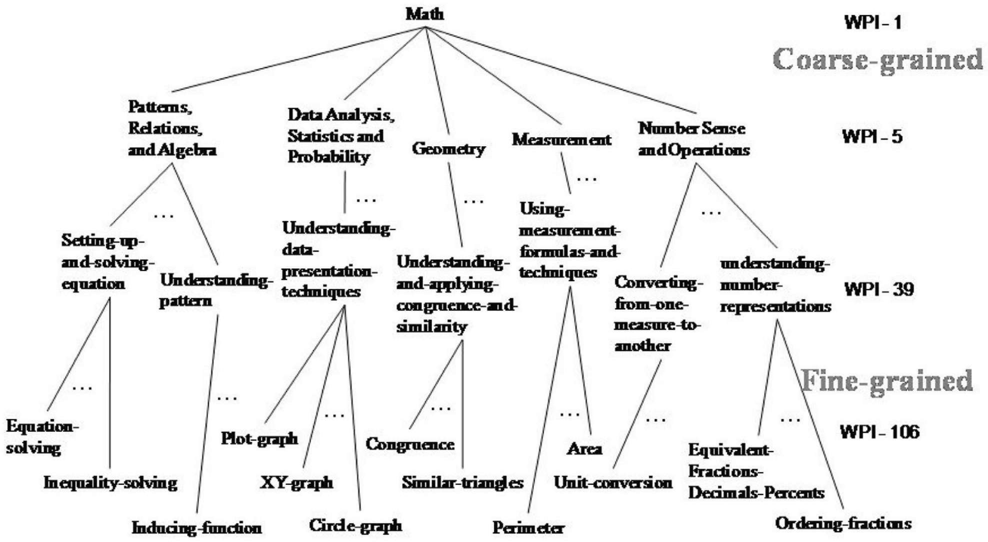
The ASSISTments approach is task-centric (development of main questions and scaffolding materials starts from released state exam items) instead of skill-centric (tasks are developed based upon the skills to be focused) and it directly attributes individual differences to unobservable skills or other latent variables. In April 2005, we invited our subject matter expert to conduct cognitive task analysis over the released state test items. They set out to make up skills and tag the entire existing 8th grade MCAS items with these skills. There were about 300 released test item to code. 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 upon our subject-matter expert that no one item would be tagged with more than 3 skills. The subject manner expert was free to make up whatever skills she thought appropriate. Although we have English names for the skills, those names are just a handy tag; the real meaning of a skill must be divined by the questions with which it is associated. We ended up with a cognitive model of 106 skills that we refer to as WPI-106.

One aspect of the project is that the ASSISTment System must serve a variety of stakeholders, and not all of them need or want reports at the same level of granularity. Therefore, even though WPI-106 may be closer to optimal for providing teacher feedback, we built other cognitive models at coarser grain size[11]. We used the fine-grained model to guide us to create the coarse-grained models. We decided to use the same 5 strands that both the National Council of Teachers of Mathematics uses, as well as the Massachusetts Department of Education. These categories 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 5 categories, but our mapping was not necessarily the same as the State's. Furthermore, we allowed multi-mapping, i.e., allowing an item to be tagged with more than one skill. In addition to the model with 5 strands (i.e. the WPI-5), a middle-level model of 39 skills was also derived (the WPI-39) by nesting a group of skills from WPI-106 into the each one of the 39 learning standards from Massachusetts Curriculum Framework. We will refer the unidimensional model as WPI-1 where "math" is the only skill in it.

Figure 6 shows the hierarchal nature of the relationship among the three models of different grain size. We think we will be able to track student knowledge in the fine- grained model, partly because students usually finished many problems along the year (on average 100+ main problems, 200+ scaffolding questions) and partly because the strategy of scaffolding questions gives us the *identifiablity*. As mentioned before, when creating scaffolding questions for the tutoring session, the authors try to focus on one skill at a time, which makes the scaffolding questions as *cognitively diagnostic assessment* on top of tutoring therefore student response to scaffolding questions can be used to determine what are the skills students have mastered or have difficulty on.

A secondary purpose of the ASSISTments builder was to aid the mapping between skills and the questions. As they are building content, the authors use the builder to tag certain skills to specific problems to indicate that a problem requires knowledge of that skill.

Figure 6. Cognitive models in a hierarchy structure


Math
WPI-1
Coarse-grained
Patterns, Relations, and Algebra
Data Analysis, Statistics and Probability
Geometry
Measurement
Number Sense and Operations
WPI-5
...
Setting-up-and-solving-equation
Understanding-pattern
Understanding-data-presentation-techniques
Understanding-and-applying-congruence-and-similarity
Using-measurement-formulas-and-techniques
Converting-from-one-measure-to-another
understanding-number-representations
WPI-39
Fine-grained
Equation-solving
Inequality-solving
Inducing-function
Plot-graph
XY-graph
Circle-graph
Congruence
Similar-triangles
Perimeter
Area
Unit-conversion
Equivalent-Fractions-Decimals-Percents
Ordering-fractions
WPI-106

Inference of Skill Level and Reporting to Teachers

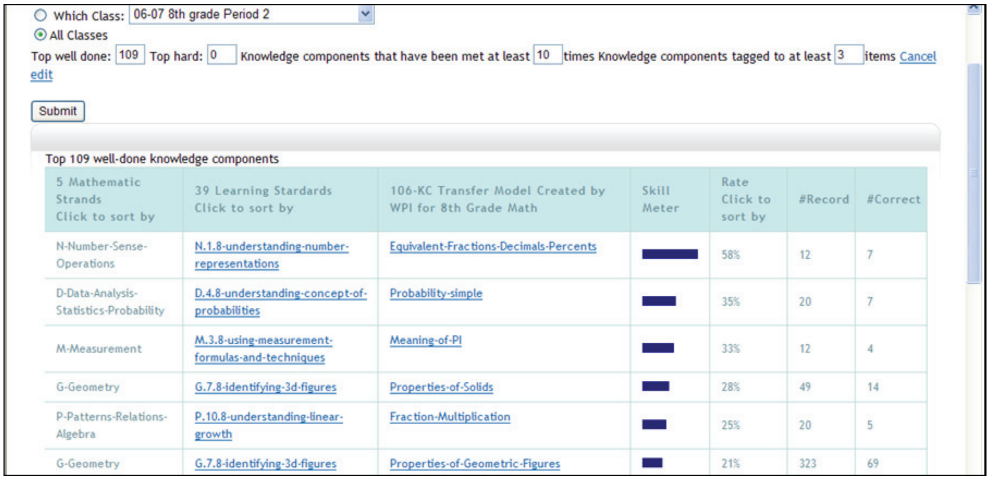
Mapping between skills and problems allows our reporting system to track student knowledge over time using longitudinal data analysis techniques (Singer & Willett, 2003; Fitzmaurice, Laird, & Ware, 2004) and make inference of student proficiency level on each skills based on their performance on the problems. In ASSISTments, the inference of student proficiency level is rather simple. Students get full credit for a skill when they correctly answer the questions tagged with the skill. In the case of a wrong answer to a question tagged with multiple skills, the system relies on response to scaffolding questions (typically tagged with only one skill) to determine which skill "to blame" (i.e., the cause of the wrong answer to the main question). If no scaffolding questions available, the most difficult skill will be blamed. Thus, connection between proficiencies and tasks is relatively loose and informal.

Turning to teacher feedback, Figure 7 shows the skill analysis report that informs teachers about the knowledge status of selected classes. Skills (labelled as Knowledge Components in the report) are ranked according to their correct rate—labelled "Rate" and displayed as "Skill Meter"—which is the percent correct rate at the items tagged with that skill model. Skills in the reports are organized hierarchically in three levels of grain sizes. Links in the column entitled "39 learning standards" will lead to the definition of a particular learning standard in Mathematics Curriculum Framework. Clicking the name of a skill in WPI-106 (labelled as "106-KC Transfer Model"), teachers will be redirected to another page showing the items tagged with that skill. 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 it. By presenting such a report, we hope we can help teachers to decide which skill and items should be focused on to maximize the gain of students' scores at a class level when instructional time is limited.

The Effect of Model Granularity on Student Performance Prediction

Model granularity is important because students seem to reason at many grain sizes, that is, students can have both deep and shallow knowledge (e.g., Koedinger & Nathan, 2004; Feng & Beck,

Figure 7. Class level skill analysis report



○ Which Class: 06-07 8th grade Period 2
● All Classes
Top well done: 109 Top hard: 0 Knowledge components that have been met at least 10 times Knowledge components tagged to at least 3 items Cancel edit

Submit

Top 109 well-done knowledge components

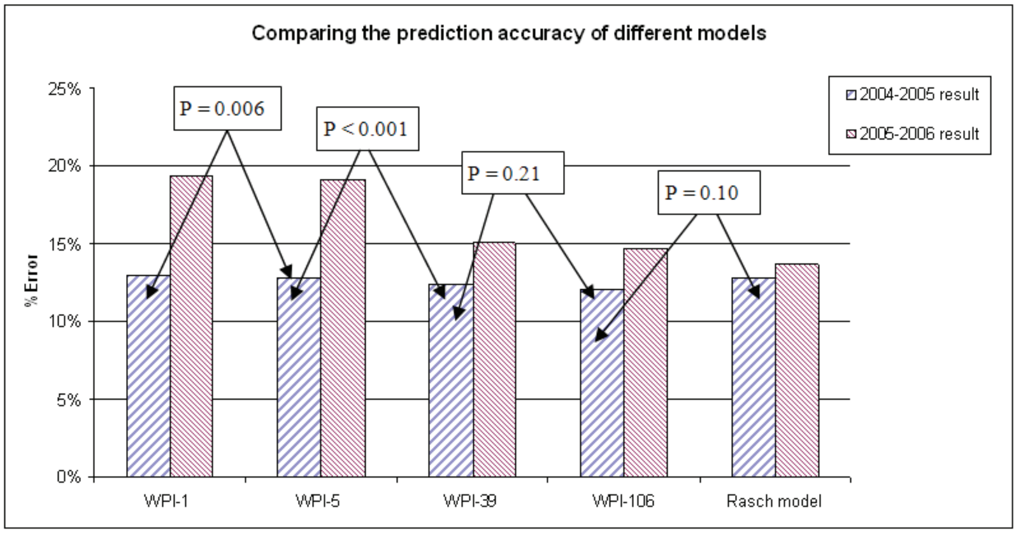
5 Mathematic Strands Click to sort by	39 Learning Stardards Click to sort by	106-KC Transfer Model Created by WPI for 8th Grade Math	Skill Meter	Rate Click to sort by	#Record	#Correct
N-Number-Sense-Operations	N.1.8-understanding-number-representations	Equivalent-Fractions-Decimals-Percents	▬	58%	12	7
D-Data-Analysis-Statistics-Probability	D.4.8-understanding-concept-of-probabilities	Probability-simple	▬	35%	20	7
M-Measurement	M.3.8-using-measurement-formulas-and-techniques	Meaning-of-PI	▬	33%	12	4
G-Geometry	G.7.8-identifying-3d-figures	Properties-of-Solids	▬	28%	49	14
P-Patterns-Relations-Algebra	P.10.8-understanding-linear-growth	Fraction-Multiplication	▬	25%	20	5
G-Geometry	G.7.8-identifying-3d-figures	Properties-of-Geometric-Figures	▬	21%	323	69

2009). The range of understanding the student has, from a deep mental model through a shallower strategy to surface code, encapsulates three simultaneously held perspectives on the problem's solution at three different grain sizes. Now that we constructed a hierarchical structure of cognitive models of various granularity, we engaged in an effort to investigate if we could do a better job of end-of-year standardized test by modeling individual skills in a finer grain size (Pardos et al., 2007; Pardos et al., 2006; Feng, Heffernan, Heffernan, & Mani, 2009) than coarser grain sizes. We considered the 4 different models: unidimensional, 5 skills, 39 skills, and 106 skills and analyzed usage data of 921 students from two school years (2004-2005, and 2005-2006). We fit mixed-effects logistic models to track student knowledge change longitudinally and then made a prediction of how a student would perform on each skill in the MCAS test[12] at the end of the year. The measure of model performance was the accuracy of the predicted MCAS test score based on the assessed skills of the students (%Error). We found out that the WPI-106 model was superior in terms of prediction accuracy of MCAS tests, followed by the model with 39 skills, and then the model with 5 skills. This suggested that finer-grained skill models were more helpful in tracking students' knowledge over time. In particular, for 2004-2005 data (shown in Figure 8.), the WPI-106 still did better than the WPI-39, but the difference was not reliable (p = 0.21). Yet, they were both statistically significantly better than the WPI-5 and WPI-1. And the WPI-106 predicted MCAS scores marginally better (p =0.10) than the Rasch model (van der Linden & Hamilton, 1997) that has been widely used in computer adaptive testing for assessment. For the 2005-2006 data, the WPI-106 model was statistically reliably better than the WPI-39, WPI-5 and WPI-1 (p<0.001 in all cases). Yet, the Rasch model produced a significant lower error than the WPI-106 model in the 2005-2006 data. We also tried to add item difficulty parameters obtained from a Rasch model into the mixed-effects logistic models as an additional covariate to account for the fact that questions tagged with the same skill may vary on difficulty, but the result suggested that the combination does not improve prediction (Feng & Heffernan, 2007a).

CONCLUSION, CONTRIBUTION & MORE DISCUSSION

In this chapter, we addressed the testing challenge in the ASSISTment system, a web-based e-learning

Figure 8. Comparing the prediction accuracy of different models on two years' data



and e-assessment system. We concentrated on the assessment ability of the system. Some evidence was presented that the online assessment system did a better job of predicting student knowledge by being able to take into consideration how much tutoring assistance was needed. Promising evidence was also found that the online system was able to track students' learning during a year well, and fine-grained skill models can be used in ASSISTments to better track student knowledge than coarse grained models. Furthermore, we showed how individual skills were modeled in the ASSISTment System and being used to give feedback to teachers. The results presented in this paper further showed that not only can reliable assessment and instructional assistance be effectively blended in a tutoring system, but more importantly, such a system can provide teachers with useful fine-grained student-level knowledge they can reflect on and adjust their pedagogy.

This chapter's main contribution lies in two aspects. First, our work provides evidence for the value of assistance measures, like the percent correct for scaffolding, the use of help, hint requests, time to respond, and other factors for prediction of standardized test scores. Traditional assessment usually focuses on whether a student answered a question 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 work, we take advantage of ASSISTments to collect extensive and rich data while students interact with the system. While the combined model leads to the best predictions, the relative success of the assistance model over the lean model highlights the power of the assistance measures. Not only is it possible to get reliable test information while "teaching on the test," data from the teaching process actually improves reliability. While the nature and amount of help that a student is given appear to be of obvious value in judging a student's mastery of knowledge, we claim that operationalizing the idea and assessing whether this promise is achievable in an implemented system is an important contribution.

Second, we demonstrated the value of a very fine-grained versus more coarse-grained model within intelligent tutoring systems. 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 et al., 2005, 2006), our 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. We are not the only ones who have been concerned with the question of granularity. Early in 1994 when speculating on long term research goals, 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 we know, our investigation is entirely unique in that it rigorously evaluates the effect of the level of details within the skill models using real student data from a tutoring system, going from very broad to very detailed. Since intelligent tutoring systems tend to use fine-grained skill models, the work validates a core underlying assumption. 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 & Murnane (2004). They developed an alternative curriculum framework and 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. Unlike us, they were using simulated users and have not applied their approach to real student data. Carmona1 et al. (2005) also introduced hierarchy structure into their Bayesian models. Yet, they focused on prerequisite relations among skills but not the various granularities of skill models, as we did.

Currently we are beginning to focus statistical modeling work on improving student knowledge modeling in the ASSISTment System. For example, Cen, Koedinger & Junker (2006) model learning curves using ideas of Draney, Pirolli & Wilson (1995). Another approach as elaborated in the Evidence Centered Design (ECD, Mislevy, Steinburg, & Almond, 2003) framework gives special attention to the role of probability-based reasoning in accumulating evidence across task performances, in terms of belief about unobservable variables that characterize the knowledge, skills, and/or abilities of students. Another approach combines the knowledge tracing algorithm of Corbett, Anderson & O'Brien (1995) with Bayes Net (DINA) models (Junker & Sijtsma, 2001). The key criterion in determining which approach to pursue is how well a model predicts data and how interpretable the model is.

One 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 system 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.

Many states are moving towards adopting "*value-added*" assessments as a new way to measure teaching and learning, which allows the states to identify not only the progress made by individual students but also how much individual teachers, schools and districts have contributed to the progress. Using this approach, researchers can predict the amount of growth those students are likely to make in a given year, based on a review of students' test score gains from previous grades. Thus, value-added assessment can tell whether a student has made expected progress, or he/she achieves beyond the expected progress. Such systems could benefit from data collected every two weeks, instead of once a year, thereby allowing schools to more quickly figure out what works at increasing student learning. Because the ASSISTment System teaches while it assesses, it makes the testing more politically palatable. In fact, this article provides evidence that because the system teaches while it assesses, it does a better job of assessing (if you hold the number of items done constant, instead of time). In the future, states and school systems may benefit from implementing an assessment tracking system that offers continuous assessment every few weeks in addition to, or even or instead of, a test that happens once a year.

Last, we want to address again that student performance modeling and assessment is not the ultimate goal, but it can provide a compass for helping us get to the ultimate goal of better student achievement. 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 a 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 in ways that people understand."[13] There is a lot of hard work to be done for teachers to change their teaching. In the education research area, new practices are being proposed to aid this change. 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 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[14]) that involves participants of teachers, graduate students, and an ASSISTment 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 2009-10 school year and a college seminar entitled *Using Advanced Educational Technology to Support Data-driven Decision Making*[15] is available for teachers. Though it is a challenging task, we believe in the power of the data. As Arne Duncan has suggested, we just need to be "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". [16]

ACKNOWLEDGMENT

This research was made possible by the U.S. Department of Education, Institute of Education Science (IES) grants, "Effective Mathematics Education Research" program grant #R305K03140 and "Making Longitudinal Web-based Assessments Give Cognitively Diagnostic Reports to Teachers, Parents, & Students while Employing Mastery learning" program grant #R305A070440, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All the opinions, findings, and conclusions expressed in this article are those of the authors, and do not reflect the views of any of the funders. We would like to acknowledge Joseph Beck for his insightful comments on this research.

REFERENCES

Ahn, J., Brusilovsky, P., Grady, J., He, D., & 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)* (pp 11-20), New York: ACM Press.

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), PAGE NUMBERS.

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. doi:10.1111/j.1745-3984.2007.00043.x

Almond, R. G., Shute, V., Underwood, J., & Zapata-Rivera, D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning*, *50*, 450–460. doi:10.1016/j.ijar.2008.04.011

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

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207. doi:10.1207/s15327809jls0402_2

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

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum: LEA.

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

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

Barnes, T. (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In Beck. J. (Ed), *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Menlo Park, CA: AAAI Press.

Beck, J. E. (2006). Using learning decomposition to analyze student fluency development. In *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (pp 21-28)*. Jhongli, Taiwan. eck, J. E., Chang, K., Mostow, J., & Corbett, A. (2008). Does help help? Introducing the Bayesian Evaluation and Assessment methodology. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* (pp 383-394).

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 meeting, Detroit, MI.

Brusilovsky, P., & 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, 4321*, 3-53. Berlin Heidelberg New York: Springer-Verlag.

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

Brusilovsky, P., Sosnovsky, S., & 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: Lecture Notes in Artificial Intelligence*, vol. 3538. Berlin: Springer Verlag

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

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,* Springer-Verlag, Berlin Heidelberg.

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.

Bull, S., & McEvoy, A. T. (2003). An Intelligent Learning Environment with an Open Learner Model for the Desktop PC and Pocket PC. In Hoppe, U., Verdejo, F., & Kay, J. (Eds.), *Artificial Intelligence in Education* (pp. 389–391). Amsterdam: IOS Press.

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 Modeling Methods that Help Learners Understand Themselves, International Conference on Intelligent Tutoring Systems*, (pp. 5-13).

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

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.). *10th International Conference on User Modeling 2005,* (pp. 347-356), Berlin, Germany: Springer-Verlag

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*. Menlo Park, CA: AAAI Press.

Cen, H., Koedinger, K., & Junker, B. (2006, June). *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.

Chang, K., Beck, J., Mostow, J., & Corbett, A. (2006, June). *A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems.* Presented at Eigth International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan

Collins, J., Greer, J., & Huang, S. (1996). Adaptive assessment of using granularity hierarchies and Bayesien nets. In *Lecture Notes in Computer Science: Proceedings of the third Intelligent Tutoring Systems*, (pp. 569 -577), London: Springer-Verlag.

Computing Research Association. (2005). Cyber infrastructure for Education and Learning for the Future: a Vision and Research Agenda. *Final report of Cyber learning Workshop Series workshops held Fall 2004 - Spring 2005 by the Computing Research Association (CRA) and the International Society of the Learning Sciences (ISLS).* Retrieved from http://www.cra.org/reports/cyberinfrastructure.pdf

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.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278. doi:10.1007/BF01099821

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

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.

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*, (pp. 148-157). Springer-Verlag, Berlin Heidelberg.

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

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.

Feng, M., Beck, J., & Heffernan, N. (2009a) Using learning decomposition to analyze instructional effectiveness in the ASSISTment system. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-2009).* Brighton, UK: IOS Press.

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.). In *Proceedings of the 1st International Conference on Education Data Mining.* Montreal, Quebec.

Feng, M., Beck, J., & Heffernan, N. T. (2009b). 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.

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), 207–230.

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, Quebec.

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*, retrieved from/doi.ieeecomputersociety.org/10.1109/TLT.2009.17.

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, Chicago.

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 Annual World Wide Web Conference.* New York: ACM Press.

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,* (pp 31–40). Springer-Verlag: Berlin.

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*.

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

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

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 www.jtla.org.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*, 75–111. doi:10.1037/0033-2909.124.1.75

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. Presented at FLAIRS2006, Florida.

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. doi:10.1177/01466210122032064

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* (pp. 17-24), Association for the Advancement of Computing in Education (AACE).

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*, Maceio, Brazil.

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.

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

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

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.* Menlo Park, CA: S. Department of Education, Office of Planning, Evaluation and Policy Development. Prepared by SRI International.

Mislevy, R. J. (1995). Probability-based reasoning in cognitive diagnosis. In Nichols, P., Chipman, S., & Brennan, R. (Eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.

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. doi:10.1007/BF01126112

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67. doi:10.1207/S15366359MEA0101_02

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*. Berlin Heidelberg, Germany: Springer-Verlag.

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

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

Olson, L. (2005). Special report: testing takes off. *Education Week, November 30*, 10–14.

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).* IOS Press. Brighton, UK.

Pardos, Z., & Heffernan, N. T. (2009b). Determining the Significance of Item Order In Randomized Problem Sets. In *Proceedings of the Second International Conference on Educational Data Mining*. Cordoba, Spain.

Pardos, Z. A., Feng, M., Heffernan, N. T., & Heffernan, C. L. (2007). Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effect Methods. In Luckin & Koedinger (Eds.). In *Proceedings of the 13th Conference on Artificial Intelligence in Education*. pp. 626-628. Amsterdam, Netherlands: IOS Press

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.

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.* Berlin, Germany: Springer-Verlag.

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

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

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.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. 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,*(pp 555–562). Amsterdam: ISO Press.

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*, (pp. 635-644. Springer-Verlag: Berlin.

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: IOS Press.

Razzaq, L., Parvarczki, J., Almeida, S. F., Vartak, M., Feng, M., Heffernan, N. T., & Koedinger, K. (2009). *The ASSISTment builder: Supporting the Life-cycle of ITS Content Creation*. IEEE Transactions on Learning Technologies.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*, 1–29. doi:10.1207/s15516709cog2101_1

Ritter, S., & Koedinger, K. (1996). An architecture for plug-in tutor agents. *Journal of Artificial Intelligence in Education*, *7*(3-4), 315–347.

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

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.

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

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and occurrence*. New York: Oxford University Press. doi:10.1093/acprof:oso/9780195152968.001.0001

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.

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

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.* (pp. 929-231). Amsterdam: IOS Press.

United States Department of Education. (2003). *Using data to influence classroom decisions*. Washington, DC. Retrieved November 2008, from http://www.ed.gov/teachers/nclbguide/datadriven.pdf van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., & Taylor, L. (2005). The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence in Education*, *15*(3), 1–47.

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 (*pp. 382-391). Springer-Verlag: Berlin.

Wiliam, D. (2006). Formative Assessment: Getting the focus right. *Educational Assessment*, *11*(3&4), 283–289. doi:10.1207/s15326977ea1103&4_7

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*, IEEE Press.

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, California.

KEY TERMS AND DEFINITIONS

ASSISTment: A collaborative project among Worcester Polytechnic Institute, Carnegie Mellon University and local public schools. The website www.assistment.org provides free accounts to teachers and students. It allows students to practice math problems as well as getting tutoring from the computer tutor. It also informs teachers of their students' progress frequently online. The term "ASSISTment" was coined by Kenneth Koedinger from Carnegie Mellon University to blend assess**ment** and **assist**ing.

Cognitive Diagnostic Assessment: Usually designed to measure student's mastery states of specific knowledge and their cognitive processing skills in a given domain. It provides detailed information about strength and weakness of students' cognitive skills. Results from cognitive diagnostic assessment have great potential to inform instruction to help student learning.

Dynamic Assessment (or Dynamic Testing): A kind of interactive testing process. During dynamic assessment, students are provided with assistance when they are having difficulty with test items. The amount of help needed is then used to supplement accuracy data (wrong/right scores) for characterizing student proficiency. Dynamic assessment allows differentiating of student capability at intermediate level between unaided success or failure.

Formative Assessment: Different from summative assessment, formative assessment refers to the assessment items or tests that are administered to students along with instruction, instead of at the end of a period of instruction. Formative assessment differs from summative assessment on the information it seeks and how the information will be used as well. It aims at providing feedback for teachers and students to adapt the teaching/learning process in order to meet learners' need.

Longitudinal Modeling: Refers to the process of tracking learner's progress over a long period of time and then building a learning trajectory by applying statistical model approaches. The longitudinal models can then be used to analyze the trend of learning progress, predict later performance, to detect impact of certain factors on learning, etc.

Student Modeling: Refers to the process of constructing student models that contains description of students' knowledge, skills or behaviors based on their responses, which can then be used to predict their performance, errors, or learning time, etc. The process of student modeling is crucial for an intelligent learning environment to be able to adapt to the needs and knowledge of individual students

ENDNOTES

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

[2] The term "ASSISTment" was coined by Kenneth Koedinger and blends Assess*ment* and *Assist*ing.

[3] The ASSISTment project is funded by grants from the U.S. Department of Education, the National Science Foundation, and the Office of Naval Research.

[4] All students in the tested grades who are educated with Massachusetts public funds MCAS tests are mandated to participate in MCAS. The Massachusetts Education Reform Law of 1993 requires that all students who are seeking to earn a high school diploma must earn certain scores on the MCAS English Language Arts (ELA) and Mathematics tests, in addition to meeting all local graduation requirements. http://www.doe.mass.edu/mcas/overview.html

[5] In the United States, students enrolled in middle school are around ages from 11 to 14, and students in high schools are usually from the age of 14 or 15 to 17 or 18.

[6] http://www.doe.mass.edu/mcas/2002/results/

[7] This does ignore the time it takes authors to plan the ASSISTment, work with their subject-matter expert, and any time spent making images and animated gifs. All of this time can be substantial, so we cannot claim to have tracked all time associated with creating content.

[8] 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, but just retrievals from short-term memory.

[9] Delayed feedback is similar to worked-out examples (Renkl, 1997) in that it shows the solution to a problem all at once. The difference is that students are administered problems first, but get no feedback until they complete a problem set. After finishing each problem, students will be told "we will give feedback at the end of the problem set."

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

[11] As the ASSISTment system is considered in multiple states and other jurisdictions, additional transfer models will be needed, that are aligned to those states' learning standards.

[12] All items in the MCAS tests are tagged in all the four models before this analysis by our subject manner expert.

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

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

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

[16] http://ies.ed.gov/director/conferences/09ies_conference/duncan_transcript.asp