

April 2012

# New Measurement Paradigms



Community for Advancing  
Discovery Research in Education

Prepared for CADRE by:

**Michael Timms—Australian Council for Educational Research**

Douglas H. Clements—University of Buffalo

Janice Gobert—Worcester Polytechnic Institute

Diane Jass Ketelhut—University of Maryland-College Park

James Lester—North Carolina State University

Debbie Denise Reese—Wheeling Jesuit University

Eric Wiebe—North Carolina State University

*April 2012*

**EDC**

**Abt**

**PSA**

**UMASS DONAHUE INSTITUTE**



This project is funded by the National Science Foundation, grant # 0822241. Any opinions, findings, and conclusions or recommendations expressed in this materials are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## CONTENTS

<b>I. Introduction .....</b>	<b>4</b>
What Do We Mean by New Measurement Paradigms? .....	5
An Overview of the Papers.....	5
<b>II. Project Descriptions.....</b>	<b>7</b>
Cognitive Tutor.....	8
Crystal Island: Outbreak .....	10
SAVE Science .....	10
Science ASSISTments.....	11
Selene: A Lunar Construction Game—A CyGaMEs Learning Environment .....	13
SimScientists Calipers II: Using Simulations to Assess Complex Science Learning .....	14
Using Rule Space and Poset-based Adaptive Testing Methodologies to Identify Ability Patterns in Early Mathematics and Create a Comprehensive Mathematics Ability Test .....	15
<b>III. Methods.....</b>	<b>17</b>
1. Knowledge Specification: A Priori Requirement for Effective Embedded Assessment .....	18
Advancing the Nation’s Instructional Technology Agenda Through Specification and Alignment .....	18
Future Work in A Priori Knowledge Specification .....	31
2. Item Response Theory and Beyond.....	33
Use of Item Response Theory and Other Measurement Methods in Intelligent Learning Environments....	33
Q-matrix Theory and the Rule Space Method.....	34
Poset Models, RSM, and Computer Adaptive Testing.....	35
Future Work .....	36
3. Machine-Learning Methods .....	37
Methods Used in Intelligent Learning Environments.....	37
Using Hidden Markov Models for Tutorial Dialogue .....	45
Future Work with Hidden Markov Models.....	46
4. Applying Educational Data Mining in e-Learning Environments.....	47
Principal Components Analysis .....	49
Pearson Correlation Coefficients.....	50
Cluster Analysis .....	51
Next Steps .....	52
<b>IV. Summary.....</b>	<b>53</b>
<b>V. References .....</b>	<b>57</b>



---

# I. INTRODUCTION

---

This collection of New Measurement Paradigms papers represents a snapshot of the variety of measurement methods in use at the time of writing across several projects funded by the National Science Foundation through its REESE and DR K–12 programs. All of the projects are developing and testing intelligent learning environments that seek to carefully measure and promote student learning, and the purpose of this collection of papers is to describe and illustrate the use of several measurement methods employed to achieve this. The papers are deliberately short because they are designed to introduce the methods in use and not to be a textbook chapter on each method.

The New Measurement Paradigms collection is designed to serve as a reference point for researchers who are working in projects that are creating e-learning environments in which there is a need to make judgments about students' levels of knowledge and skills, or for those interested in this but who have not yet delved into these methods.

## What Do We Mean by New Measurement Paradigms?

The New Measurement Paradigms group formed around a common interest of a diverse group of researchers who all felt the need to discuss and explore new ways of thinking about the field of educational measurement. Researchers in the group have backgrounds in the disciplines of computer science, educational measurement, cognitive science, instructional technology, mathematics education, statistics, psychology, and learning and teaching, which reflect our belief that the type of measurement of student learning required in complex learning environments transcends traditional demarcations between disciplines.

As the New Measurement Paradigms name implies, we feel it is time for a new conception of what constitutes the field of educational measurement. The field of educational psychometrics grew up around the types of responses that were typical in paper-based, large-scale assessments: primarily multiple-choice and written-response items. Over the years, the methods of analyzing these types of student responses have become increasingly sophisticated, progressing from Classical Test Theory to Item Response Modeling methods that can model different dimensions of students responses and even dynamically adapt the assessment to the ability of the student as the student is being assessed. But computer-based learning environments create new demands of the measurement methods by increasing the ways in which a student can respond during an assessment task and by requiring that assessments be conducted “on-the-fly” so that assistance can be provided to learners as they progress. As a result, the old measurement models that were built to cope with full sets of response data at the end of a set of tasks are no longer adequate on their own.

In the approach advocated by the New Measurement Paradigms group, methods for educational measurement must be expanded by adopting, adapting, and mashing methodologies from a range of disciplines. As individuals working in the development and testing of electronic learning environments for the future, we all need to become familiar with a wider range of methods than we have used in the past. This group of papers is a step along that road.

## An Overview of the Papers

To put the papers presented here into a context that might be familiar to others in the field of education, it is useful to frame them in terms of the Conceptual Assessment Framework presented by Mislevy, Almond, and Lukas (2003) in their introduction to *Evidence Centered Design*. That Conceptual Assessment Framework defines five models: (1) Student Model, (2) Evidence Model, (3) Task Model, (4) Assembly Model and (5) Presentation Model. The papers in this collection address the Student Model and the Evidence Model parts of the framework. The Student Model is the representation of the knowledge,



skills, and abilities that are to be measured and the Evidence Model details how the direct observations of student actions are processed and interpreted.

The Student Model is addressed in two of the New Measurement Paradigms papers. The first paper, *Knowledge Specification: A Priori Requirement for Effective Embedded Assessment*, addresses a key aspect of designing the student model by exploring how knowledge representation is handled in three different types of learning environment. The fourth paper, *Applying Educational Data Mining in E-learning Environments*, discusses how methods from the emerging field of educational data mining can be applied to explore the datasets that result from students' actions in complex learning tasks in order to determine patterns of learning that can help developers reshape the tasks and design the measurement models needed to interpret the students' actions.

The second paper, *Item Response Theory and Beyond*, and the third paper, *Machine Learning Methods*, both address the Evidence Model part of the Conceptual Assessment Framework, which deals with how data are analyzed and interpreted. While the papers discuss distinct measurement methods (such as Item Response Theory, Bayes Nets, and Hidden Markov Models), they share some common features. The first common feature is that they all model variables that represent features of the students' knowledge state that cannot be directly observed. This leads to another commonality—that each of the systems is thus making inferences about those states of learning based on collecting data on things that can be observed, such as a student's response to a question or his or her action during a task. This, of course, happens in traditional assessments, but unlike those assessments in which there is a complete set of data at the end of a test, in the learning environments featured in these papers, measurement takes place throughout the learning tasks and so decisions are made when there are only small quantities of evidence. This, in turn, leads to yet another commonality—that several of the methods are probability based because there is a fair degree of uncertainty in the measurement. Other commonalities among the Item Response Theory, Bayes Nets, and Hidden Markov Models described in the third and fourth papers are that each of the models requires training based on data or on the opinion of experts in the domain being taught. They all also share the need for having some measures of model fit so that judgments can be made about whether the selected model and its design is doing a good job of modeling the data from student responses and actions.

The final section of this collection of papers looks forward to where research in the field of educational measurement in electronic learning environments is headed and identifies areas that still need further research and development.

We hope you will be inspired by the set of papers and will seek to learn more about the methods described here.

---

## **II. PROJECT DESCRIPTIONS**

---



This section contains brief descriptions of the projects within which the measurement methods discussed in the Methods section are used. The project descriptions are presented in alphabetical order and include:

- Cognitive Tutors – Carnegie Mellon University
- Crystal Island: Outbreak – North Carolina State University
- SAVE Science – University of Maryland (College Park)/Temple University
- Science ASSISTments – Worcester Polytechnic Institute
- Selene – Wheeling Jesuit University
- SimScientists – WestEd
- Using Rule Space/Post-based Adaptive Testing – University of Buffalo

## Cognitive Tutors

Carnegie Mellon University

Principal Investigator: Ken Koedinger

Grant numbers: ARI 90385K0343<sup>1</sup>; ARI 90389K0190; NSF-DRL 8470337; NSF-IIS 8318629; NSF-PHY 8715890; NSF-DRL 8954745; NSF-DRL 9253161; ONR N0001484K0064<sup>2</sup>; ONR N00014870103; ONR N0001491J1597

Cognitive Tutors are a form of educational technology that takes a primarily problem-based or learning-by-doing approach and is able to adapt instruction to individual student needs. The adaptation is made possible through cognitive psychology research, which is used to identify the components of knowledge (e.g., skills and concepts) that students need for successful performance in demanding contexts, and artificial intelligence technology, which is used to implement these knowledge components in a *cognitive model*—a computer simulation of the correct and common incorrect ways students (try to) solve problems (or task scenarios) in the domain of interest.

Figure 1 shows a screen shot from a unit within the *Cognitive Tutor Algebra* course. This course is in regular use, about two days a week, by over a half million students a year.

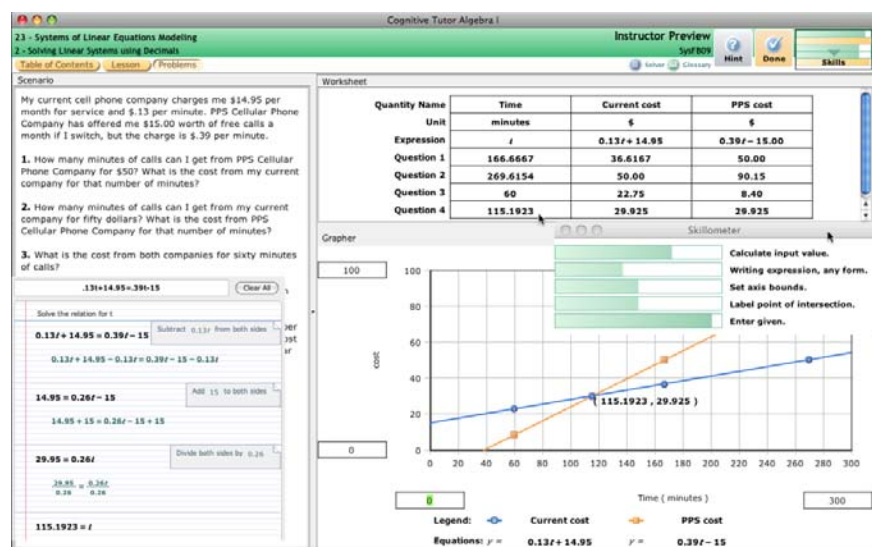


Figure 1. Screenshot from the *Cognitive Tutor Algebra* course. Used with permission.

<sup>1</sup> Army Research Institute

<sup>2</sup> Office of Naval Research



Cognitive Tutors provide two key forms of adaptation. The first involves providing students with instruction (guidance or assistance) in the context of learning by doing, typically while solving problems in a science, technology, engineering, or math domain. This instruction is adapted to the individual student because different students may have different strategies for solving problems and may struggle in different places within a problem. To provide instruction adapted to each student's solution progress, Cognitive Tutors use an algorithm called *model tracing*, which uses the cognitive model to keep track of where in a problem solution a student is. Thus, to be sure, when given challenging problems, novice students experience difficulties at different points. At any such point, the Tutor can compute, from the cognitive model, what are the reasonable next steps. Thus, if the student's next step is not reasonable (e.g., it is a common error), the tutor can provide immediate feedback and keep the student from wasting time and getting unnecessarily frustrated. Further, if the student is stuck and asks for help, the Tutor uses the model to generate a hint for a reasonable next step for the student to pursue. In this way, students only get instruction when they need it, and then are back on their own to discover useful concepts and strategies and to strengthen their skills in problem solving.

The second key form of adaptation comes between problem-solving tasks and involves selecting a next task for a student that best meets the students' learning needs. An algorithm called *knowledge tracing* keeps track of the evolving growth of students' knowledge, including what skills and concepts they have mastered and which need more work. Cognitive Tutors use the results of knowledge tracing to pick a next task for a student that addresses skills or concepts they have not yet mastered. In this way, students can focus their learning efforts more optimally, with different students getting different problems, some moving more quickly, and others getting more help as needed.

The cognitive model drives the effectiveness of Cognitive Tutors, not only in supporting model and knowledge tracing, but also in helping the developer design appropriate task sequences that are system to the learning trajectories of different students. A better cognitive model means a more effective tutor, and research has demonstrated that getting the cognitive model right (through cognitive task analysis) is non-trivial. For example, Koedinger and Nathan (2004) found that, in contrast to the beliefs of math educators, beginning algebra students are better able to solve story problems than corresponding equations. The cognitive model in the CTA was based on such research and thus the Tutor works differently than it would without such a data-driven design.

Cognitive Tutors have been subject to many experimental evaluations both in the laboratory and in the field. One study of college students' use of Tutor for programming demonstrated students learned more *and* in one third the time than students learning programming with typical instruction. Large-scale, full-year classroom evaluations of the *Cognitive Tutor Algebra* course have been published in peer review journals (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley & Mark, 1997; Ritter, Anderson, Cytrynowicz & Medvedeva, 1998), and met the strict What Works Clearinghouse standards (e.g., WWC, 2009), and have demonstrated better student learning when compared to a business-as-usual algebra course. However, some evaluations have found no difference in student learning (e.g., WWC, 2010). Without going into the many potential reasons for different outcomes (e.g., classroom implementation differences, insensitivity of assessments, curriculum differences), it is clear that this approach can be improved and better outcomes can be achieved.



## **Crystal Island: Outbreak**

*North Carolina State University*

*Investigators: James Lester (PI), John Nietfeld, Hiller Spires*

*Grant numbers: NSF-IIS 0812291; NSF-DRL 0822200; NSF-DRL 1138497*

Recent years have seen a growing recognition of the importance and challenges of creating learning environments that promote motivating, inquiry-based science learning. Pedagogical agents are embodied software agents that have emerged as a promising vehicle for promoting effective learning. They provide customized problem-solving experiences and advice that are precisely tailored to individual learners in specific contexts. By co-habiting a rich inquiry-based learning environment with learners, pedagogical agents can meticulously observe learners' problem solving activities, offer situated advice, and actively support learners' iterating through cycles of questioning, hypothesis generation, data collection, and hypothesis testing. However, inquiry-based learning also presents a significant challenge: the very "openness" of the learning environment introduces multiple sources of complexity into tutorial planning. To address the complexities associated with scaffolding inquiry-based learning, this project explores Bayesian pedagogical agents that leverage recent advances in Bayesian and decision-theoretic computational models of reasoning to promote self-regulated learning experiences that are both effective and engaging.

The project has two complementary technology and learning thrusts:

It will develop a full suite of Bayesian pedagogical agent technologies for inquiry-based science learning environments. To promote effective and engaging learning processes and outcomes, the research team is creating Bayesian pedagogical agents that leverage probabilistic computational models that systematically reason about the multitude of factors that bear on decision making to infer learners' beliefs, goals, and plans, including strategy use, from their problem-solving actions. By introducing pedagogical agents into the visually engaging environments that typify high-end game platforms and embedding them in dynamically generated science narratives, we are addressing the complementary goals of achievement and engagement.

It will provide a comprehensive account of the cognitive processes and results of interacting with Bayesian pedagogical agents in inquiry-based science learning by conducting extensive empirical studies. To understand the cognitive mechanisms by which self-regulated inquiry-based science learning occurs with middle school students interacting with Bayesian pedagogical agents, the research team is taking a multi-method approach to investigating the use and effectiveness of Bayesian pedagogical agents. In both controlled laboratory and classroom-based field settings, these studies are investigating the central issues of self-regulation with respect to both achievement (science content knowledge, transfer, and effective strategy use, including strategy selection and strategy shifting) and engagement (self-efficacy, situational interest, and mastery orientation with an emphasis on persistence) to determine precisely which technologies and conditions contribute most effectively to learning processes and outcomes.

## **SAVE Science**

*University of Maryland, College Park/ Temple University*

*Investigators: Diane Jass Ketelhut (PI)*

*Grant numbers: NSF-DRL 0822308; NSF-DRL 1157534*

Situated Assessment using Virtual Environments for Science Content and Inquiry (SAVE Science) is an innovative project that explores the use of immersive virtual environments (IVEs) for assessing both science content and inquiry in middle schools. A series of game-like situated assessment modules (or

quests) are designed around curricula taught in middle school science classrooms. The quests are targeted at specific standards that are currently assessed via district-wide multiple-choice based tests. It is hypothesized that this new way of assessment will help us better capture and analyze students' understanding in science.

Students gather data useful in solving the assessment problem through interacting with nonplayer characters (NPCs), exploring and observing visual and tacit clues, and applying virtual tools—"sci-tools"—for varying purposes like measuring, graphing, and note taking. The quests open with a non-player character posing the problem. There are additional NPCs present in the world in different quests that provide clues or distracters to the students. All student actions in the world are recorded in an underlying database and can be used to map and understand student knowledge.

Currently, the project has designed three assessment modules on topics of: weather patterns, beginning speciation, and gas laws. It is hypothesized that student choices in solving the inquiry-based problems in our modules, as indicated by actions undertaken in the modules, can provide insights into student understanding of the content being assessed. In addition, two conditions for success are examined: how best to design the games to account for and manage the cognitive load students experience while conducting inquiry in complex virtual environments, and how to help teachers integrate technology into their pedagogy. The project is also examining how these assessments change teacher and student perceptions of efficacy and behaviors. By providing detailed feedback to teachers on their students' performance, it is possible to change current teacher practices. The broader impact of this study is the attempt to contribute to research using VEs to assess student performance in science.

SAVE Science is currently being implemented in schools in multiple public school districts in the Mid-Atlantic region of the United States. Across all of our modules over the last two years, approximately 2,000 students (representing diverse racial, ethnic, and free lunch status) have taken one or more of our assessments.

Project researchers are particularly focusing on three student variables: student performance in solving the assessment problem in the SAVE Science modules, cognitive load (mental effort), and self-efficacy.

We gather data to investigate these variables by:

- Capturing each student's actions and pathways in the world in an underlying database
- Surveying students on their views of overall mental effort required by the modules, and their self-efficacy in science, scientific inquiry and technology use
- Asking students about their gender, ethnicity, SES, and prior technological experience
- Interviewing students and teachers about their experiences
- Conducting classroom-based discussion on their solutions to the problem

## Science ASSISTments

*Worcester Polytechnic Institute & Carnegie Mellon University*

*Investigators: Janice D. Gobert (PI), Ryan Baker, Neil Heffernan, Joseph Beck & Ken Koedinger*

*Grant numbers: NSF-DRL 0733286; NSF-DGE 0742503; NSF-DRL 1008649; U.S. Dept. of Ed. R305A090170*

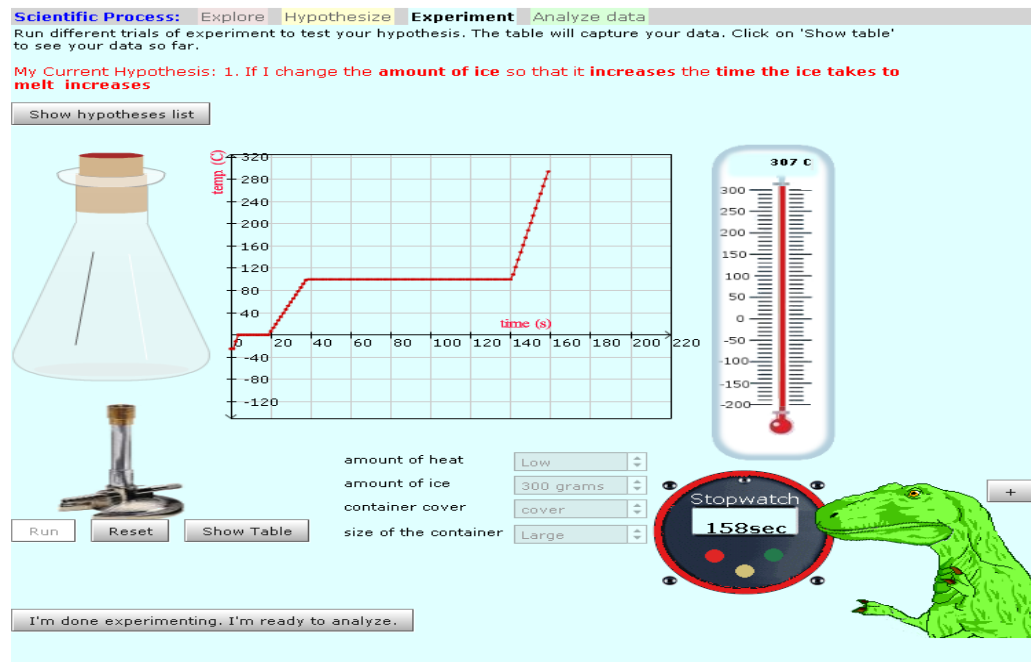
Science ASSISTments (Assessing Students' Inquiry Skills in Real Time with the Goal of Scaffolding Students in Real Time, (<http://www.scienceassistments.org>) is a learning and assessment environment for physical, life, and Earth science that assesses middle school students' scientific inquiry skills, namely, generating hypotheses, conducting experiments, interpreting data, and warranting claims with evidence. In addition to our 25+ microworlds across the three science domains, the project developed a suite of



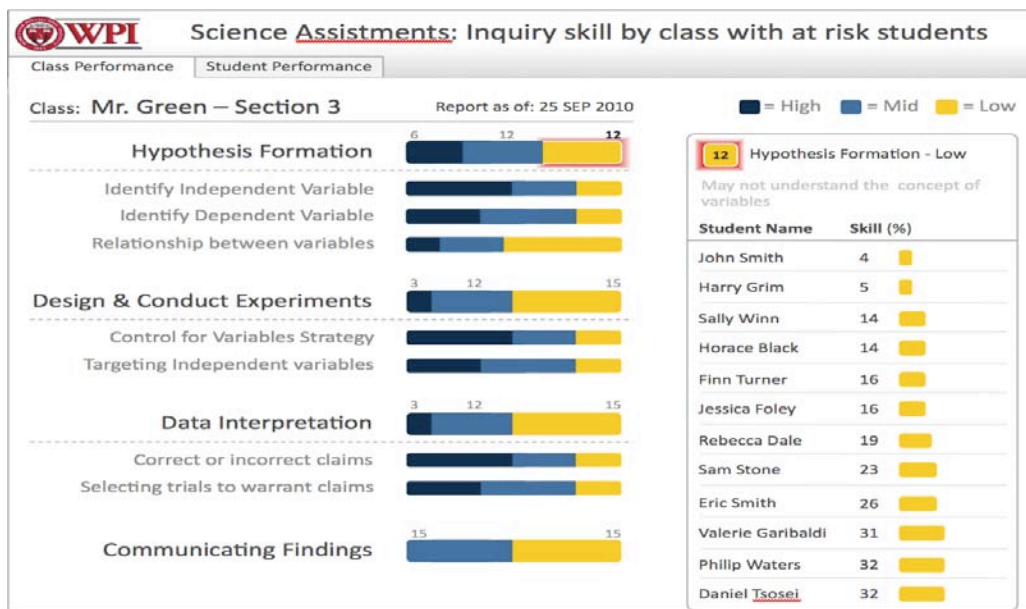
inquiry tools to support students' inquiry on the four skills mentioned above; project researchers also designed a reporting system for teachers. Together, our microworlds, the logging functionality, and the inquiry tools both provide the basis for conducting performance assessment in real time and has the provisions for adaptive scaffolding of students' inquiry in real time. The reports provide teachers with performance assessment of their students' skills, allowing teachers to make curricular decisions in real time.

For well-defined inquiry skills, we autoscore students' inquiry using knowledge-engineering algorithms; for more complex inquiry skills, we use algorithms generated via educational data mining. Respectively, these were developed using production rules, and a combination of text replay tagging and machine learning. By reacting to students' inquiry strategies in real time, it is hypothesized that the project will positively affect both students' science process skills and students' content learning. Specifically, researchers measure students' inquiry skills in terms of improvement at: generating testable hypotheses, testing their articulated hypotheses (as opposed to testing other, random hypotheses), conducting controlled experiments, correctly interpreting data, and warranting their claims with appropriate data. Secondly, researchers measure content knowledge gains using pre- and post-test items from standardized-type content assessments.

Currently, project researchers are testing the system with regard to its efficacy at honing students' inquiry skills in real time by conducting a series of randomized controlled studies in our partner schools; the demographics of these students represent a wide range of SES and ethnic backgrounds, and thus, the data should generalize well.



**Figure 2.** Screen shot of Science ASSISTments environment with state change microworld. (The students' hypothesis is given at the top; Rex, our pedagogical agent, is pictured on the right.)



**Figure 3.** Inquiry assessment report from Science ASSISTments system. (Data by class are on the left and data for each student are on the right.)

## Selene: A Lunar Construction Game—A CyGaMEs Learning Environment

Wheeling Jesuit University

Investigators: Debbie Denise Reese (PI), Charles A. Wood and Ben A. Hitt

Grant number: NSF-DRL 0814512



Imagine sitting at your computer and blasting away at what will quickly become a full-fledged pockmarked moon like our own. All the while you're learning about the solar system's basic geological processes AND helping test videogames as learning tools. That's what learners have in store when they play *Selene: A Lunar Construction Game*. *Selene* players construct the Earth's Moon, then replicate its 4.5-billion-year



history as they pepper it with impact craters and flood it with lava flows. *Selene* aligns with the new STEM education framework (Committee on Conceptual Framework for the New K-12 Science Education Standards & National Research Council, 2011):

- *Selene* players discover and apply foundational Earth and planetary science concepts.
- *Selene* gameplay is goal-seeking behavior through which players model physical science and cross-cutting concepts.
- Players must analyze and interpret data to progress and win.

*Selene* is a single-player Flash/Flex game available 24/7 via the Internet: <http://selene.cet.edu>.

*Selene* is also a research environment designed to investigate causal claims through random assignment, manipulation of conditions, and embedded assessment of player behavior and affect. Although the game



targets players ages 9 and up, learners ranging from first grade through graduate students have played. The top 10 game scores post from players age 9, 10, early tweens, and undergraduates. The leading player is 9 years old. Game completion takes 2 to 10 or more hours, although some fifth-graders completed in 45 minutes. Players from formal, informal, and nonformal contexts across the United States and internationally have registered to play. In compliance with federal requirements, all *Selene* youth players first submit parent/guardian permission through an approved adult volunteer (e.g., teacher, parent, neighbor, club leader, etc.).



**CyGaMEs.** *Selene* is a CyGaMEs learning environment. CyGaMEs (Cyberlearning through Game-based, Metaphor Enhanced Learning Objects) is a principled approach to instructional game design and embedded assessment. CyGaMEs designs to make learning more intuitive through goal-driven gameplay that motivates and rewards players to discover and apply targeted concepts. The key is a strong theoretical foundation in cognitive science, entertainment game design, instructional design, flow theory, and learning science. The CyGaMEs method applies analogical reasoning theory to translate domain knowledge – *what people think* – into procedural gameplay – *what learners do*. CyGaMEs specifies the target domain, designs a game world that is its relational analog, guides and constrains gameplay through game goal analogs to targeted learning goals, and measures learning as the trace of behavioral transactions toward the game goal. CyGaMEs is specific to instructional games that modify behavior to cause and measure learning.

**Learning.** CyGaMEs collects measures of learning at the grain size of player or game gesture interaction that modifies the game world. Every 10 seconds of gameplay, *Selene* posts a ratio-level Timed Report of learner behavior, measuring learner progress toward the game goal. Timed Report is a sensitive measure, accurately discriminating pre- vs. post-learning for exemplar players,  $F(1,20) = 358.73, p < .001, \text{partial } \eta^2 = .95$ , and for the general data sample,  $F(1,102) = 74.38, p < .001, \text{partial } \eta^2 = 0.42$ . Current work uses statistical regression and calculus to investigate the rate of individual player learning (slope) at every Timed Report post and acceleration in rate of learning. *Selene's* external assessments are more traditional measures that provide convergent evidence that Timed Report measures content learning and not just learning to play the game,  $F(2,13) = 18.62, p < .001, \text{partial } \eta^2 = 0.74$ .

**Affect.** The CyGaMEs flowmeter measures self-reported flow dimensions within every five-minute segment at a randomly selected, predetermined time. Replicated results build a case for the reliability of the measure for nine channels of experience: flow, arousal, anxiety, worry, apathy, boredom, routine, expertise, and control.

## SimScientists Calipers II: Using Simulations to Assess Complex Science Learning

WestEd

Investigators: Edys Quellmalz (PI), Michael Timms, Barbara Buckley, and Mark Loveland

Grant Number: NSF-DRL 1020264

One component of the WestEd SimScientists program is Calipers. The Calipers II project has developed science simulations to be used during a unit to monitor and improve student learning and also to administer as summative unit test. The simulations allow students to conduct inquiry on science phenomena that are difficult to observe directly. The science content tested is integrated into a framework of three levels of all science systems: (1) components and their roles, (2) interactions among components, and (3) emergent behaviors of systems. The simulation-based assessments test science standards seldom or only partially measured in traditional district or state on-demand, annual tests (see <http://www.simsScientists.org>).

Project goals are to: (1) Develop formative and benchmark simulation-based assessments of the system knowledge and science inquiry strategies for fundamental life and physical science topics; (2) Develop formative assessment simulation modules with immediate, individualized feedback, graduated coaching, and reflection activities; and (3) Study the feasibility, usability, utility, and technical quality of the new assessments.

The assessment development and validation involves external review, cognitive labs, feasibility testing, and pilot testing. The curriculum-embedded, simulation-based assessments provide immediate, individualized feedback and graduated levels of coaching, followed by a classroom reflection activity. Student responses are a mixture of selected responses and constructed responses such as text, arrows, sliders, and saved trials. The simulation generates a progress report for students and teachers. For the benchmark, a Bayes Net generates proficiency levels for assessment targets. A pre/post-test was developed from AAAS calibrated items.

Evidence-centered assessment design was employed to specify and link the *student model* (specifying the science knowledge and inquiry practices targets) to the *task models* (for eliciting evidence of the targets) to the *evidence model* (specifying how the observable events would be scored and reported as proficiency levels). Cognitive labs conducted with 28 students examined usability and construct validity—that the tasks and items elicited the intended science target. A field test in progress shows significant differences for the Ecosystems benchmark assessment favoring students who used the embedded simulations in comparison with students who did not.

In a related study, Integrating Simulations into Balanced State Science Assessment Systems, the Calipers Ecosystem and Force and Motion assessment suites were administered by 58 teachers in three states and 28 districts to 5,800 students to study the assessments' feasibility, utility, and technical quality. These field tests found that the Calipers II assessments promoted high levels of student engagement, had high technical quality, were feasible to implement across a diverse range of school technical infrastructures, and were considered instructionally useful. Students performed better on the benchmark simulation assessments than on the traditional post-test items, and performance gaps of English language learners (ELL) and students with disabilities (SWD) compared to other students were reduced (Quellmalz, Timms, Silbgerglitt, & Buckley, 2012).

A policy brief developed in the three-state study recommended models for integrating science simulations into balanced state science assessment systems. The curriculum-embedded assessments can serve formative, classroom purposes. Calipers II benchmark assessments can also serve as summative assessment components for curriculum program evaluations and for research projects.

## Using Rule Space and Poset-based Adaptive Testing Methodologies to Identify Ability Patterns in Early Mathematics and Create a Comprehensive Mathematics Ability Test

University at Buffalo, Case Western Reserve University, Columbia University Investigators: Douglas H. Clements (PI), Julie Sarama, Curtis Tatsuoka, and Kikumi Tatsuoka  
Grant Number: NSF DRL-1019925

In the present project, the research team is conducting cognitive and statistical analyses of an existing research-based mathematics assessment using the theories and procedures of Kikumi and Curtis Tatsuoka, including Q-Matrix theory, the Rule Space Method (K. K. Tatsuoka, 2009), and poset-based adaptive testing methodologies (C. Tatsuoka & Ferguson, 2003).



The project has two major objectives. First, it will produce a reduced and adaptive assessment that will take less than one-half the time to administer and yet will yield more useful and detailed information about children's knowledge of mathematics, including their level of thinking along multiple empirically validated developmental progressions and detailed cognitive profiles. Second, in accomplishing this, researchers will evaluate, refine, and elaborate these cognitive developmental progressions; operationally define the cognitive attributes (e.g., concepts and skills) that constitute each level of thinking in those progressions; and empirically evaluate the theoretical model that includes these attributes.

The extant version of the Research-based Early Math Assessment (REMA) measures core mathematical abilities of children from age 3 to 9 years using an individual interview format, with standardized protocol and scoring procedures (Clements, Sarama, & Liu, 2008; Clements, Sarama, & Wolfe, 2011). Abilities are assessed according to theoretically and empirically based developmental progressions that underlie the research-based learning trajectories. We are creating a reduced measure, the Comprehensive Research-based Early Math Ability Test (CREMAT), which will use Computer Adaptive Testing (CAT) to guide the dynamic selection of items, and score and analyze the results, providing a comprehensive report including the overall score, level of achievement within each developmental progression (and, if requested, detailed report on achievement on each item in that topic), and a cognitive profile.

Given the existence of both the present REMA and a large database of results of testing, the development and the evaluation of the new instrument will be interwoven, iterative processes. Project researchers are (a) analyzing the knowledge and skill required by each item on the REMA, initially guided by the research-based developmental progressions on which they were based (create a Q-Matrix); (b) evaluating the veracity of these through an examination of the extant scores and procedural ("strategy") coding of the REMA assessments and by follow-up analyses of the video to check and elaborate these analyses as needed (construct a rule space by expressing all of the knowledge states as the points in the rule space and classifying students using legacy REMA student response data sets); (c) applying statistical methods to analyze the instrument and identify essential attributes (map knowledge states into the Rule Space classification space and also fit poset models); (d) evaluating and validating cognitive models through data analysis and comparison to video and coding; (e) creating a reduced and adaptive measure that will provide more information about children in dramatically shorter time period (redesign item pool and construct adaptive test); and (e) evaluating and validating this measure (administer to sample, collect qualitative and quantitative data; analyze classification properties).



---

## III. METHODS

---



# 1. Knowledge Specification: A Priori Requirement for Effective Embedded Assessment

- *Debbie Denise Reese—Wheeling Jesuit University*
- *Janice Gobert—Worcester Polytechnic Institute*

“Show me!”

That’s what learners and educators should demand of learning environments: show what students already know, what students need to learn, how students learn, what they have learned, and when they have learned it. Showing what students learn is the job of formative and summative assessment, and this job is important. Decades of educational research confirm that aligned and effective assessment can enhance teaching and learning (e.g., Bransford, Brown, & Cocking, 2000; Pellegrino, Chudowsky, & Glaser, 2001). Learners and educators may one day demand “Show me now!” because in formative assessment, sooner is better than later. They may also demand stealth assessment; that is, authentic, performance-based, and embedded assessment tools that measure knowledge and learning as they are demonstrated by real-time learner performance. Situating and embedding assessment within authentic activity diverge from more traditional assessments that cloister test-taking and decontextualized assessment instruments and measures. Cyberlearning technologies may advance progress toward meeting those demands. In fact, the nation’s cyberlearning research and implementation agenda targets embedded assessment as a way to enhanced teaching and learning (Borgman et al., 2008; National Research Council [NRC], 2011; U.S. Department of Education Office of Educational Technology, 2010).

Cyberlearning technologies support authentic assessment when they embed data production and collection within interactivity. Such interactivity is a [virtual] physical transaction along a continuum of situatedness. For example, learners may select a step in a cognitive tutor’s problem-solving process, adjust levers within a simulation, or enact gameplay that changes the state of a game world. Cyberlearning technologies are already data-driven systems. Collection of learner transactions simply requires posting of learner transactions as data. However, data are just that: data. They are value neutral. Data may carry noise or signals. Assessment of learner behavior as a measure of knowledge acquisition and application requires that those data carry a signal. A signal is analyzed for information that can be interpreted and reported. Some researchers and developers borrow a bottom-up approach from qualitative research (for grounded theory overview, see Corbin & Strauss, 1990; Strauss & Corbin, 1991, 1994) and interpret emergent data patterns for stories of learners’ knowledge acquisition and application. Other researchers follow a top-down approach. They pre-specify and align targeted knowledge with assessment information, signal characteristics, data, learner gestures, and the instructional system. Both approaches require a priori analysis to align the learning environments and embedded assessment with targeted knowledge. This section characterizes a selection of research and development approaches as top-down applications of a priori knowledge specification. These approaches implement as three cyberlearning technologies: cognitive tutors, simulations, and instructional games.

## Advancing the Nation’s Instructional Technology Agenda through Specification and Alignment

Each of these three technologies—cognitive tutors, simulations, and instructional games—has the capacity to enhance teaching and learning through instruction and integrated (embedded) assessment. Game-based instruction with aligned assessment is the youngest of the three (National Research Council, 2011). However, the technologies overlap, and the boundaries among them are fuzzy. Only a modicum of imagination conjures a future in which effective instructional environments seamlessly morph and mix



attributes and affordances of cognitive tutors, simulations, and instructional games. Viable cyberlearning instruction and assessment require specification of targeted knowledge and alignment among the to-be-learned knowledge, the cyberlearning environment, and the assessment (Clark, Feldon, Merriënboer, Yates, & Early, 2008). Lack of specification can lead to inaccurate or incomplete instruction, causing learner misconceptions. Misalignment and missing alignment among learning goals/domains, learning environment, and embedded measures produce unserviceable assessment. In other words, effective instructional design of cyberlearning environments requires (a) formal a priori specification of to-be-learned knowledge and (b) its alignment with instruction and assessment. Cognitive tutors—like the algebra, geometry, and LISP Cognitive Tutor environments produced at Carnegie Mellon University (CMU)<sup>3</sup> (Koedinger, & Corbett, 2006)—are the most mature of the three types of environments. The CMU Cognitive Tutor derives from a formal cognitive task analysis that specifies an expert model. This and subsequent research determine and refine a domain-specific student model that drives the tutor’s instructional decisions, such as scope, sequence, pacing, and feedback (Stamper & Koedinger, 2011).

Review of published research suggests that state-of-the-fields for instructional simulations and instructional games lag behind the older sibling (Quellmalz, Timms, & Schneider, 2009). Quellmalz et al. write that “. . . in practice, many simulation-based learning programs fail to identify the specific knowledge and skills that are targeted. . . .” (2012, p. 57). For example, Quellmalz, Timms, and Schneider reviewed 79 articles that investigated the use of simulations in grades 6–12 and included reports of measured learning outcomes (National Research Council, 2011, p. 88). Overall, those reports lacked sufficient detail about domain specification and alignment among the three components (the to-be-learned knowledge, instruction, and assessment) to draw or support conclusions about students’ science knowledge and process skills (Quellmalz et al., 2009). Quellmalz et al. observed that game-based embedded assessment is in its “infancy” (p. 11) and that effective instructional gaming requires (a) a design phase that aligns the game with the to-be-learned and (b) embedded assessment.

This paper discusses examples of NSF-supported projects that follow principled approaches to knowledge specification and alignment: (1) Carnegie Mellon’s Cognitive Tutors, (2) Worcester Polytechnic Institute’s Science ASSISTments, (3) WestEd’s implementation of evidence-centered design principles (ECD: Mislevy, Almond, & Lukas, 2003) in *SimScientists* simulation-based assessments (Quellmalz, Silberglitt, & Timms, 2011) and (4) the CyGaMEs (Cyberlearning through Game-based, Metaphor Enhanced learning objects) approach to instructional game design and embedded assessment as implemented in *Selene: A Lunar Construction Game* (Reese, 2009; Reese et al., in press).

The CyGaMEs approach is specific to instructional game-based (that is, goal-centric) technologies that modify behavior to cause and measure learning. Like a *SimScientists* simulation, CyGaMEs environments contain simulations that permit the learner to manipulate “. . . structures and patterns that otherwise might not be visible or even conceivable . . .” (Quellmalz et al., in press). Some scholars taking a cognitive science approach to learning, including development of Cognitive Tutors, hypothesize knowledge components as “an *acquired* unit of cognitive function or structure that can be inferred from performance on a set of related tasks” (Koedinger, Corbett, & Perfetti, 2012). Like a Cognitive Tutor, CyGaMEs studies a type of knowledge component. CyGaMEs traces learning trajectories at the level of learner gestures, and learners discover and apply knowledge components.

Like Cognitive Tutors, the Science ASSISTments group has broken down inquiry strands (NRC, 1996) into sub-skills in order to score students’ inquiry processes. To do so, the group uses methods that include both knowledge engineering described later in this paper, as well as educational data mining (Sao Pedro,

---

<sup>3</sup> This section focuses on design and application for assessment. Although discussion of efficacy and scalability research studying CMU Cognitive Tutors falls outside this scope, the interested reader may refer to the What Works Clearinghouse Intervention Report (2010).



Baker, Gobert, Montalvo, & Nakama, in press; Gobert, Sao Pedro, Baker, Toto, & Montalvo, accepted), described in the paper on data mining. The Science ASSISTments project uses microworlds with which students conduct virtual inquiry.

This section overviews how design and implementation of Cognitive Tutors, Science ASSISTments, and *SimScientists* prepare to address diagnostic assessment. A final example uses *Selene* to illustrate how a CyGAMES approach conducts these processes.

## Use in Intelligent Learning Environments

### *Cognitive Tutors*

CMU's Cognitive Tutors are an application of Anderson's ACT-R (Adaptive Control of Thought—Rational, Anderson & Lebiere, 1998), which evolved from earlier, related cognitive architectures and theories of human cognition within a lineage of seminal work that traces back to Simon and Newell (Newell, 1990; Newell & Simon, 1972) and birthed artificial intelligence. ACT-R models “overt, observable human behavior” (Anderson & Lebiere, 1998, p. 10). Anderson and his colleagues classify cognitions into declarative knowledge (the *what* of knowing), propositional knowledge (the *how* of knowing), and goal structures (the *why* motivating cognitions). Procedural knowledge is comprised of condition-action units called *production rules*. Declarative knowledge chunks components of subsumed knowledge to form networks of connected components. As knowledge matures toward expertise, goals and subgoals coactivate rich networks of declarative knowledge and execute production rules (Anderson & Schunn, 2000). CMU scholars developed their Cognitive Tutors to “show that cognitive models can lead to successful learning” (Anderson, Corbett, Koedinger, & Pelletier, 1995, p. 191); that is, they used the tutors to test theories of human cognition through emulation. Research from the mid-1980s to 1990s supported efficacy, motivating the scientists to transition Cognitive Tutors from the laboratory and into educational practice. At that time, research had demonstrated the success of Cognitive Tutors (e.g., [a] efficiency: students achieved gains in one-third the time and [b] effectiveness: when time is held constant, Cognitive Tutors produced gains of one standard deviation, Anderson & Schunn, 2000), but their use had not integrated into classroom practice (Anderson et al., 1995). To this end, CMU partnered with the Pittsburgh Public Schools system to engineer Cognitive Tutor learning environments that met the district's needs, with the district in the role of client. Today, “more than 500,000 (U.S.) students per year” (Stamper & Koedinger, 2011, p. 354) use Cognitive Tutors for mathematics.

**Knowledge specification.** Scientists must conduct formal cognitive task analysis (for overview of task analysis and methods, see Chipman, Schraagen, & Shalin, 2000; Clark et al., 2008) to construct a Cognitive Tutor from a cognitive model (Corbett & Anderson, 1995). The design process begins with a cognitive task analysis to specify “overt observable behavior and cognitive functions behind it” for the domain-specific knowledge, thought processes, and goal structures underlying expertise (Chipman et al., 2000, p. 3). In general, an analyst selects applicable techniques to match targeted learning outcomes. The process might begin with a foundational literature review and progress through observations and interviews, process tracing, and/or more formal techniques that specify formal models of knowledge components. Cognitive task analysis is essential for effective instruction (Anderson & Schunn, 2000), whether that instruction implements as a Cognitive Tutor or other type of learning environment. A review by Clark et al. identified a sequence of five common steps performed in most dominant cognitive task analyses (2008, p. 580):

1. Collect preliminary knowledge.
2. Identify knowledge representations.
3. Apply focused knowledge elicitation methods.
4. Analyze and verify data acquired.
5. Format results for the intended application.

**Model and knowledge tracing.** Today, CMU scholars design Cognitive Tutors to cause and trace knowledge at the level of the knowledge component, identified as “a generalization of any element of a knowledge representation including a production rule, schema, or constraint” (Stamper & Koedinger, 2011, p. 354). For each Cognitive Tutor, initial task analysis steps determine a candidate expert model. “Cognitive modeling is expensive” (p. 203), and the CMU researchers estimate that on average they require 10 hours or more to develop each production rule. Although Cognitive Tutors are designed to advance knowledge acquisition and help novices achieve expertise, students neither think nor solve problems like experts (Koedinger & Corbett, 2006). Therefore, the expert model is refined and realized within the CMU Cognitive Tutor as a hypothetical, ideal student model representing student competence as a production set modeling the targeted knowledge set within appropriate curricular objectives for accurate modeling of learner interactions (Anderson et al., 1995). Then, CMU researchers develop and apply statistical algorithms to refine knowledge components (Stamper & Koedinger, 2011) and improve the student model. Results are used to redesign tutor sequencing, tasks, instructional communication (messages, scaffolds, and feedback), and knowledge tracing.

Knowledge tracing is used at the individual learner level for formative assessment and to structure the environments for mastery learning (Corbett, Anderson, & O’Brien, 1993). Model tracing compares the student model to individual student actions and drives the tutor’s instructional decisions, such as scope, sequence, pacing, and feedback. For example, the LISP ACT Programming Tutor (Corbett & Anderson, 1995, p. 256) applies model tracing and knowledge tracing to customize information and user interaction within interface windows that provide instruction, problem statements, exercises, hints, a menu of operator templates, an input field, and a skill meter.

If the student action matches an applicable rule, the tutor accepts the action and fires the rule to update the code window and update the internal representation of the problem state. If the student action does not match the action of any applicable rule in the ideal model, the action does not register in the code window, and the tutor provides a brief message in the hint window. . . . The skill meter . . . displays the tutor’s model of the student’s knowledge state (Corbett & Anderson, 1995, p. 256).

The skill meter represents each applicable rule by a bar graph. “Shading represents the probability that a student knows [a] rule,” and a checkmark indicates mastery. Computationally, knowledge tracing uses Bayesian algorithms (Corbett & Anderson, 1995) to estimate the probability that each knowledge component is in a learned state. Individualized and immediate assessment couples with decades of analysis results to enable these tutors to diagnose a student’s error and respond with advice and remediation (National Research Council, 2001). Over time, offline analysis and modeling of interaction logs enable the developers to analyze the system, refine the student model, refine instruction, refine problem development and selection, and/or provide feedback (Stamper & Koedinger, 2011). Indeed, returning to its *raison d’être*, cognitive tutor research has motivated and informed revision to ACT theory and architecture.

### *Science ASSISTments: Microworlds designed to assess and assist*

The Science ASSISTments group (<http://www.scienceassistments.org>) has developed an educational learning and assessment environment as well as 25+ microworld-based simulations for assessment in the domains of physical science (NSF-DRL 0733286), life science, and Earth science (U.S. Dept. of Ed. R305A090170). Students conduct inquiry with microworlds, which share many features with real apparatus (Gobert, 2005); as such, these present authentic performance assessment opportunities because the microworlds are instrumented so students can generate a hypothesis, test it, interpret data, warrant claims with data, and communicate findings (NRC, 1996; NRC, 2011). Having several microworlds allows many opportunities to assess students’ inquiry skills (Shavelson, Wiley, & Ruiz-Primo., 1999) in the context in which they are developing (Fadel, Honey, & Pasnick, 2007; Mislevy et al., 2002). The



goals of the project are to (1) provide formative and summative assessment reports about students' inquiry skills to teachers so that they can adjust their instruction in real time if they wish, (2) provide real-time feedback to students as they conduct inquiry (some implementation studies have already been completed and others are currently underway), and (3) characterize inquiry skills as they develop both within domains and across domains. The assessment algorithms are based on knowledge engineering (Feigenbaum & McCorduck, 1983; Studer, Benjamins, & Fensel, 1998) and educational data mining (Baker & Yacef, 2009; Romero & Ventura, 2007) of students' log files (Gobert et al., accepted; Sao Pedro et al., in press; Montalvo et al, 2010).

In this subsection, work which leverages knowledge-engineering, similar to Cognitive Tutors, is briefly described. In addition to using knowledge engineering, the group has successfully shown that educational data mining can be used to do automatic assessment of some key inquiry skills (Sao Pedro et al., in press; Gobert et al., accepted); this work is described in the paper on data mining in this report.

In terms of knowledge engineering, Gobert and Koedinger (2011; under review) developed as a proof of concept automated assessment of scientific inquiry skills. In particular, they used model-tracing (Koedinger & Corbett, 2006) to develop a cognitive model of science inquiry skills, particularly, testing stated hypotheses, use of the control for variables strategy, and making warranted interpretations from data. Specifically, the control for variables strategy involves the procedural and conceptual understanding of how, when, and why a controlled experiment should be conducted so that one can make valid inferences about the effects of one independent variable on a dependent variable (Chen & Klahr, 1999; Kuhn, 2005). In addition, this model provides a rich qualitative, process-oriented scoring of students' inquiry "moves" within a guided scientific inquiry simulation for the domain of state change. They addressed the validity of this automated approach to performance assessment both quantitatively, in terms of reliability and predictive validity, and qualitatively, in terms of providing rich traces of student inquiry steps and "mis-steps" or haphazard inquiry (Buckley, Gobert, Horwitz, & O'Dwyer, 2010).

In order to develop production rules to auto-score science inquiry skills, a necessary first step was breaking down the scientific inquiry skills of interest into sub-skills so that production rules could be written. Related rules were proposed in Koedinger, Suthers, & Forbus (1999), and the Science ASSISTments group has continued to work towards this goal as part of the development of the Science ASSISTments learning and assessment environment. The production rules developed in the project were designed to score domain-general aspects of scientific inquiry; that is, they can score when a student is or is not testing their hypothesis, when a student is or is not using the control for variables strategy, and when a student is making a claim based on data. The production rules are applied to domain-specific declarative facts, and thus, adaptation of only a few such declarative structures is necessary to apply the production rule system to another science domain. Such work is currently in progress (U.S. Dept. of Ed. R305A090170). See Table 1.1 below for the production rules.

**Table 1.1.** Production rules used to score inquiry skills of interest.

*Test-target-variable rule*

- IF the goal is to explore the effects of changing variable <IndVar> and you have a baseline trial  
THEN create a comparison trial by changing the value of variable <IndVar>.

*Test-nontarget-variable-error rule*

- IF the goal is to explore the effects of changing variable <IndVar> and you have a baseline trial  
THEN create a comparison trial by changing variables other than <IndVar>.

*Change-one-variable rule*

- IF the goal is to test a hypothesis and you have a baseline trial  
THEN change one variable value to create a comparison trial.

*Change-more-than-one-variable-error rule*

- IF the goal is to test a hypothesis and you have a baseline trial  
THEN change the values of two or more variables to create a comparison trial.

*Make-a-warranted-claim rule*

- IF the goal is to argue for hypothesis <H1> and there is a pair of trials that change only the independent variable and the result of those trials is consistent with <H1>  
THEN enter hypothesis <H1> in interpretation phase of the simulation interface.

Much work has been done with cognitive tutoring in domains where it has been relatively easier to develop task analyses, but less work has been done in domains that are more ill-defined, like applying science inquiry skills. Gobert and Koedinger showed that model-tracing (Koedinger & Corbett, 2006) can be used for performance assessment for science inquiry, which builds on the extensive cognitive tutor research in other better-understood domains, such as math (Koedinger & Corbett, 2006) and computer programming (Corbett & Anderson, 1995), as well as earlier work by Koedinger et al. (1999), which described a vision for a system and a prototype implementation in which students could conduct inquiry and receive tutoring.

This work also contributes to the assessment of inquiry skills and to technical methods to auto-score some critical science inquiry skills. It is an advance because to date there has been difficulty in measuring inquiry due to the amount of data required for reliable measurement (Shavelson, Wiley, & Ruiz-Primo, 1999). The approach is also novel in separating inquiry from the domain-specific knowledge (Gobert, Pallant, & Daniels, 2010; Mislevey et al., 2003), that is, the model can provide data in order to differentiate content knowledge from process skills (e.g., a correct hypothesis indicates content knowledge, but a correct analysis may not indicate prior content knowledge in that the student may have just discovered the content knowledge through their experimental trials). This capability also has the affordance of identifying interesting inquiry patterns. For example, genuine discovery is one such pattern; this may occur when a student is observed (as they have been) making a scientifically inaccurate hypothesis initially and then, through inquiry, coming to understand the effects of one variable on another and entering a correct (and necessarily different) interpretation that is warranted by relevant controlled experimental trials. Secondly, the method can also identify cases of confirmation bias, whereby the learner does not discard a hypothesis based on negative results (Dunbar, 1993; Klahr & Dunbar, 1988; Klayman & Ha, 1987; Quinn & Alessi, 1994). Lastly, because the model can conduct a fine-grained trace of students' inquiry processes, it can model different states of the learner's knowledge and skill level as they are acquired. For example, one targeted inquiry skill is controlling for variables (CVS). A skill



trajectory for CVS traces progress from starting state (a) when students first conduct trials that are neither not using CVS nor are they targeting the proper independent variable (i.e., the student is not testing their hypothesis), to instances demonstrating skill acquisition and application (b) as the student conducts successive trials, they begin to target the proper independent variable and then begin to use the control for variables strategy. This pattern has been observed in the data.

### *SimScientists Simulation-based Assessments*

WestEd's work (Quellmalz et al., 2011) leverages simulation technologies as dynamic and interactive tools for increased authenticity and accuracy in formative assessment (gives feedback to student and educator during instruction to improve student performance) and summative assessment (evaluates performance after the conclusion of instruction). The goal is to enhance capacity for assessment to measure student ability and growth in science practices of model-making required for scientific habits of mind and systems thinking in support of the 2011 framework for K–12 science education and anticipated aligned state standards (for framework, see Committee on Conceptual Framework for the New K–12 Science Education Standards & National Research Council, 2011). WestEd's *SimScientists* assessment project also applies a Bayes Net for psychometric analysis to determine student proficiency and provide feedback as learners investigate and solve simulations (Quellmalz et al., in press; see also the paper on Machine Learning).

However, for the purposes of the discussion within this section, the salient distinction between the Cognitive Tutors approach and that employed by *SimScientists* is not idiosyncratic characteristics of the psychometrics. Rather, it concerns how focus on model-based learning (here, the use of simulations to cause, enhance, and measure viable mental models for targeted science systems in life, physical, and Earth science), project aspirations, and evidence-centered design methods shaped WestEd's approach to task analysis and the grain size of the student model. Cognitive Tutors were developed to model and study cognitive architecture. *SimScientists* was developed to study the potential and efficacy of an alternative assessment that might capture evidence of deep “understanding of dynamic science systems and uses of science inquiry practice” (Quellmalz et al., 2011, p. 2). WestEd applied an evidence-centered design (ECD) approach to assessment that aligns three models: (a) student model: the model of student knowledge, skills, or other attributes to be assessed; (b) evidence model: the behaviors or performances that reveal the student model; and (c) task model: the tasks that elicit the evidence model (National Research Council, 2011, p. 89). Their approach differs slightly from the assessment triangle proposed a decade earlier by the Committee on the Foundations of Assessment, which also consists of three aligned components: (a) cognition: theory or beliefs about how students represent knowledge and develop domain- or task-specific competency, (b) observation: tasks or situations that demonstrate degree competence, and (c) interpretation: the methods and tools used to interpret evidence (observations) and make inferences about the student's competency within the targeted realm (NRC, 2001, pp. 44–49). The WestEd approach began with literature review components:

- What have the learning sciences told us about learning by this age group and in this domain? What has been studied and with what research paradigms?
- What are existing frameworks and standards<sup>4</sup>?
- Curriculum analysis: What science knowledge is represented in practice as valued for assessment and instructional targets?




WestEd's elegant and parsimonious student model for *SimScientists* ecosystems (see Figure 1.1) illustrates the high-level conceptual analysis accomplished by WestEd's ECD approach. It has identified components and roles, interactions, and emergent behaviors as generic aspects of any system. *SimScientists* contains an ecosystem module and a force and motion module. Each of these modules

---

<sup>4</sup> For example, WestEd drew from the *National Science Education Standards*, American Association for the Advancement of Science *Benchmarks for Science Literacy*, and National Assessment of Science (NAEP) 2009 framework items and results.



instantiates in parallel but discrete contexts. For example, the *SimScientists* ecosystem simulation may contextualize as a tundra, mountain lake, savanna, or grassland. A module's contexts are relational analogs; that is, each of the ecosystems functions according to the same components, behaviors, interactions, and emergent behavior. They are relational analogs. Only the superficial details (context) change. The approach described in the next subsection, CyGaMEs, uses a game world as the relational analog of a targeted domain's relational structure.

Model Levels: Generic		Content Targets	Inquiry Targets
 <p>Components and Roles</p>	<p>What are the components and behaviors of the system (at this level)? What are the "rules" of the system in general?</p>	<p>Every ecosystem has a similar pattern of organization with respect to the roles (producers, consumers, and decomposers) that organisms play in the movement of energy and matter through the system.</p>	<p>Use principles to identify role of organisms.</p>
 <p>Interactions</p>	<p>How do the interactions influence the individual components?</p>	<p>Matter and energy flow through the ecosystem as individual organisms interact with each other. Food web diagrams indicate the feeding relationships among organisms in an ecosystem.</p>	<p>Observe interactions among organisms.</p>
 <p>Emergent Behaviors</p>	<p>What is the overall state of the system that result from many interactions following specific rules?</p>	<p>Interactions between organisms and between organisms and the ecosystem's nonliving features cause the populations of the different organisms to change over time.</p>	<p>predict observe explain investigate</p>

**Figure 1.1.** *SimScientists* content and inquiry target models for ecosystems in middle school. © 2010 WestEd. Used with permission.

Again, the major shaping factors in *SimScientists* design and development were ECD and model-based learning. The WestEd developers matched assessment and instruction with science practice, concentrating efforts to assess only what could not be done by traditional paper-based assessments. They represented dynamic science systems in action to match the assessment to authentic practice of science as inquiry:

Students complete tasks, such as making observations, running trials in an experiment, recording data, interpreting data, making predictions, and explaining results. They answer questions by various methods, such as selecting from a choice of responses, changing the values of variables in the simulation, drawing arrows to represent forces, and typing explanations (Quellmalz, et al., 2011).

The assessments serve as both (a) curriculum supplements that provided immediate formative assessment and (b) summative, end-of-unit "benchmark" assessments that provided no scaffolding (feedback or coaching). The *SimScientists* system uses Bayes Nets in the benchmark assessments to analyze both the data it has collected and teacher-scored data, and it reports results to educator and student concerning (a) student achievement related to state standards, (b) specific content, and (c) inquiry practice.

The *SimScientists* assessment system has been tested in laboratory think alouds (28 middle school students), in classroom feasibility trials (two classrooms), and via a large-scale pilot study (55 teachers, 5,465 students in three states, 28 districts, and 39 schools). Testing studied usability, feasibility, and utility for formative and summative assessment (Quellmalz, Timms, Silberglitt, & Buckley, 2012), for example (summarized from the report):

- *Early usability testing* identified areas for revision to increase clarity and ease of implementation. They also indicated that *SimScientists* tasks measured targeted skills.



- *Classroom-level feasibility trials* demonstrated the assessments could successfully administer within a typical class period of 50 minutes. They also informed revisions to the simulation-based assessments, connected activities, observation, and interview protocols.
- *Large-scale pilot study* investigated four research questions: (a) reliability and validity, (b) effectiveness with English language learners and students with disabilities, (c) implementation feasibility across classroom contexts, and (d) usefulness to teachers as feedback for evaluating and adapting instruction. Overall, teachers were satisfied with the simulations and agreed that instant feedback, interactions, and visuals afforded by *SimScientists* were more beneficial than traditional paper-and-pencil assessments. Technically, the pilot demonstrated acceptable reliability across assessment items and supported validity of the simulation's classification of learners' scaffolding and coaching needs, concurrent validity because of the correlation between simulation and more traditional assessment, and discriminant validity that the simulations more effectively measured learners' inquiry practices than did the traditional assessment. The simulations reduced performance gaps for English language learners and students with disabilities, especially for inquiry skills. Results provided evidence that *SimScientists* simulations could successfully serve as assessment instruments within a state's assessment system.

WestEd *SimScientists* simulations and animations design, development (content, interactivity, scaffolds, and feedback), and analysis most probably required information derived from a finer grain of knowledge than the models, content targets, and inquiry targets illustrated in Figure 1.1. This analysis would form a specification subsumed by the overarching targets and models. For example, subsumed specification could take the form of causal chains. If so, a coherent representation of that knowledge might begin to resemble those specified and deployed by developers of Cognitive Tutors.

### *CyGaMEs Selene: A Formalism for Instructional Game Design with Embedded Assessment*

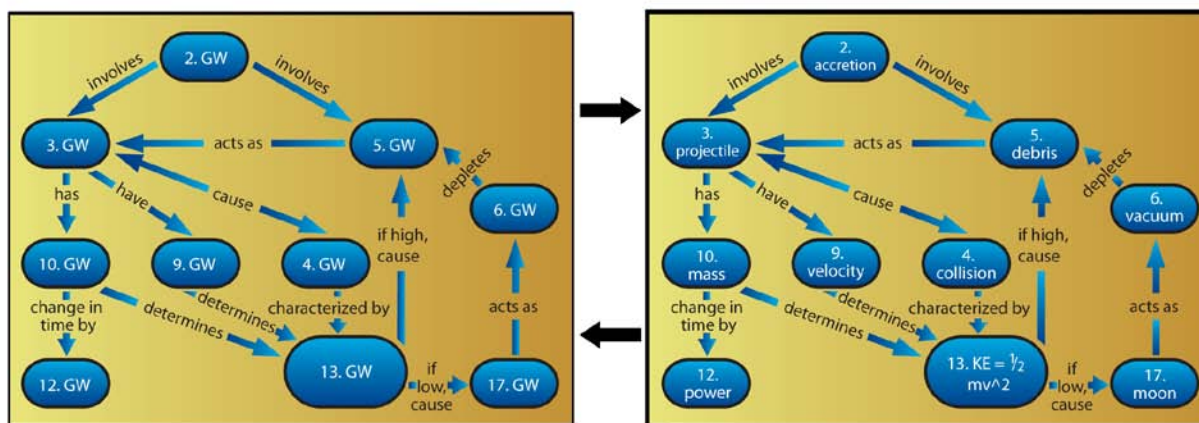
Although games and play are intrinsic components of human (Huizinga, 1950) and even animal (e.g., the play of seals, dogs, bears, deer) culture, the technologies serving goal-centric virtual environments designed for [virtually] embodied gameplay are a relatively recent invention and innovation. Will Wright is a recognized and awarded leader and one of the most successful and influential designers of interactive entertainment<sup>5</sup>. He observed that the field of *entertainment* game design is young and *instructional* game design is even younger (Langhoff et al., 2009). He said that both fields lack principled approaches to design practice. Wright said the need for principled design motivates his efforts to disseminate his game design methods through presentations at events like the Game Developers Conference (personal communication, January, 26, 2008; for examples of Wright's presentations about his game design practice, see Wright, 2003, 2006a, 2006b, 2009). Wright noted that design principles have yet to consistently filter into formal entertainment game design instruction. Where entertainment games privilege fun and marketability, instructional games must privilege the accuracy of alignment to targeted content. CyGaMEs provides one formal approach to instructional game design with embedded assessment. CyGaMEs is specific to instructional games that modify behavior to cause and assess learning, although CyGaMEs can be more widely applied.

CyGaMEs also specializes in assisting novices to construct viable prior knowledge necessary for future knowledge acquisition (Anderson, Reder, & Simon, 1998; Merrill, 2002). The CyGaMEs approach applies analogical reasoning theory (specifically structure mapping, see Gentner, 1983; pragmatic constraints, see, for example, Holyoak & Thagard, 1989) to direct and constrain mapping from a specified knowledge domain to a relationally analogous game world (see Figure 1.2 and Reese, 2009). People naturally learn from interacting with their everyday world and experience (Figure 1.2, left rectangle; relations are the directional arcs) by making inferences from the known, concrete, or familiar to the

---

<sup>5</sup> E.g., 2001 Lifetime achievement award, Game Developers Choice Awards, 2002: Academy of Interactive Arts & Sciences Hall of Fame, 2005 PC Magazine Lifetime Achievement Award

unknown or abstract (right-hand rectangle). This cognitive process is analogical reasoning. Learning scientists specify the relations that connect domain components. This specification is signified by the right-hand rectangle. In CyGaMEs, a domain is specified in the right-hand rectangle, and then a game world is designed to map to it: (a) the knowledge domain specification translates into the game world, gameplay mechanic, and game system, and (b) the learning goal(s) translates into the game goal and subgoals, guiding and reinforcing player discovery and application of targeted concepts. The CyGaMEs approach translates abstract concepts experts *think* about natural phenomena into concrete, procedural gameplay that players discover and *do*. CyGaMEs design commences with required cognitive task analysis that results in a knowledge specification akin to an ontology. To date, CyGaMEs specifications have been produced as concept maps, but many types of knowledge representations will suffice if they represent systems of concepts and the relations between them (Gentner, 1980).



**GW = Game World**

**Figure 1.2.** An excerpt from the CyGaMEs *Selene* domain specification, illustrating the mappings from real world (left) to concept (right) and from conceptual domain (right) to game world analog. © 2009 Debbie Denise Reese and Charles A. Wood. Used with permission.

**Selene.** CyGaMEs *Selene* is a game about the origin and evolution of the Earth’s Moon. The cognitive task analysis produced a *Selene* specification of 101 related domain concepts and took more than a month of expert interviews and analysis. Players discover and apply planetary geology and physical science concepts as they select and gather materials to form and differentiate the Earth’s Moon, then pepper it with impact craters and flood it with lava flows. Conceptually, *Selene* aligns with the new K–12 science framework (Committee on Conceptual Framework for the New K–12 Science Education Standards & National Research Council, 2011):

- (1) Scientific Practices (1, 2, 3, 4, and 5).
- (2) Cross-Cutting Concepts (1, 2, 3, 4, 6, and 7).
- (3) Disciplinary Core Ideas Physical Sciences (PS 2.B, PS2C, PS3.A, PS3.B., PS3.C., PS4.C<sup>6</sup> (MoonGazer postgame unit activities), ESS1.B, ESS1.C<sup>7</sup>).

Thus, the *Selene* environment’s application of the CyGaMEs approach is as relevant for accomplishing the nation’s education goals as it is for cyberlearning research investigating (a) how to design instructional games and (b) how to embed and interpret assessment. *Selene* game modules are interactive simulations of physical phenomena. The game provides formative assessment (feedback to player) of

<sup>6</sup> MoonGazer postgame unit activities only.

<sup>7</sup> Primary objective. The framework recognizes that scientists study bodies such as the Moon to provide information about Earth’s formation and early history.



accomplishment at the gesture level (the grain at which player interaction with the system changes the state of the game world) and the module level (change in content focus). Proposed future work includes a system to automate and report analyses of individual and aggregate player learning and affect to players and their educators and researchers.

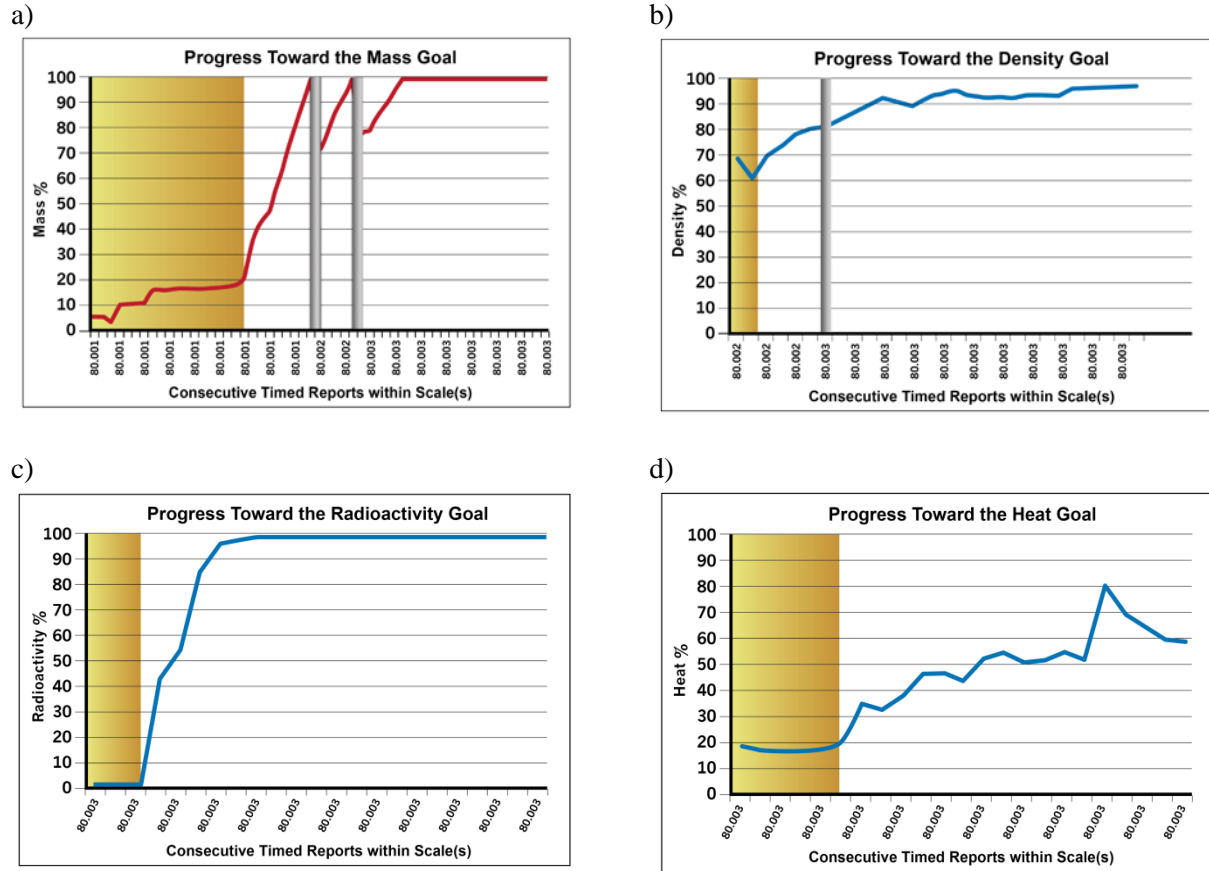
**Timed Report: A CyGaMEs Assessment Suite Tool.** By design, CyGaMEs player progress toward game goal(s) is a measure of targeted preconceptual or conceptual mental model growth. This is the Timed Report post. CyGaMEs posts player activity at the gesture level, and each gesture posts with its parameters and identifiers. For example, the slingshot gesture used by players to accrete the Moon posts with the following parameters and identifiers: speed, direction, projectile radius, accuracy, density, game scale, game module, timestamp in milliseconds, server time, and player ID. Every 10 seconds, *Selene* evaluates gestures (gameplay) for progress toward the game goal. In addition to identifiers like timestamp and player identification, the post contains baseline state, goal, and current progress at the culmination of the time segment. “Timed Reports provide a meaningful synopsis of player behavior above the gameplay gesture level” (Reese et al., in press).

The 10-second interval was selected for practical and theoretical reasons. Practically, it considers the time required to initiate and complete a *Selene* gameplay gesture (about one to three seconds) and drain on computer resources. Theoretically, the 10-second interval derives from Allan Newell’s (1990) powers of 10 analysis of the time scale of human action. Ten seconds is the level of the unit task, the longest of levels within the cognitive band. Newell wrote, “This is where knowledge is available from the environment about what to learn, namely experience attained in working at ~ 10 sec,” and it is governed by “the opportunity for acquisition” (Newell, 1990, p. 149). Its cause locates in “conditions for learning” and works “according to representational law. The processes have been so composed that they represent things—objects, relations, and activities.” This level supports the next, the rational level, which is goal directed.

The CyGaMEs team initially established Timed Report validity through triangulation of video data capturing gameplay activity with velocity gestures (Reese & Tabachnick, 2010). Then, hand and algorithmically analyzed Timed Report data for unidimensional goals successfully categorized pre versus post learning for exemplars (Reese & Tabachnick, 2010) and across the general corpus of player data (Reese et al., in press). CyGaMEs has identified learning moments and learner states (e.g., progressing, failing) algorithmically through comparisons of slope and variance calculated over consecutive running windows (see Reese et al., in press, for description of procedure). Using the Timed Report, the CyGaMEs team identified learning moments, pre versus post learning gameplay, and investigated the interplay between affect (for description of the CyGaMEs flowometer, see Reese, 2010) and learning. Once the learning moment is identified and data are categorized as pre versus post learning moment, subsequent analyses employ standard statistical techniques, such as multilevel modeling and within-between ANOVAs. Plots of learning traces over time evidence targeted concept formation and application while demonstrating that each player’s path and gameplay are idiosyncratic. Statistically, hierarchical linear modeling has shown that individual differences are no longer significant once a learning moment predictor (data within participant is categorized into pre versus post learning) is entered into the model (Reese & Tabachnick, 2010).

CyGaMEs Timed Report research expanded to investigate multidimensional goal states. Figures 1.3a–3d illustrate player progress for four *Selene* accretion module learning goals over three scales of gameplay using the slingshot gesture and core gameplay mechanic. *Selene* measures scale 1 learning as the unidimensional mass goal (players learn the principle of accretion to build the Moon’s mass by adjusting the size of particles selected and velocity of the slingshot gesture). Vertical bars in Figures 1.3a and 1.3b indicate scale divisions. In scale 2, players strive toward multidimensional goals; mass and density objectives are concurrent. In scale 3, players work toward four concurrent goals: mass, density,

radioactivity, and heat. The graphs in Figure 1.3 plot percent of player progress toward each goal. Trace over time is flat (slope  $\sim 0$ ) or negative (slope  $< 0$ ) until a learning moment. Overall, a player's post-learning-moment progress plots with a positive slope or evens out once the player has reached a goal ceiling.



**Figure 1.3(a-d).** Trace of player progress toward learning goal over time. Percent goal accomplishment for one player on 55 consecutive timed reports (posted every 10 seconds of gameplay) for four *Selene* Accretion module learning goals over three scales of increasing challenges. Ordinal axis indicates progression in time from left to right, game module (80 = accretion), and the scale (the thousandths-place digit: 1, 2, or 3). Gold shading indicates pre-learning moment gameplay. Vertical bars for Figures 1.3a and 1.3b indicate change of scale. Goal 1.3a increases at scale change, reducing player accomplishment to 70% at the start of both scale 2 and scale 3. © 2011 Debbie Denise Reese. Used with permission.

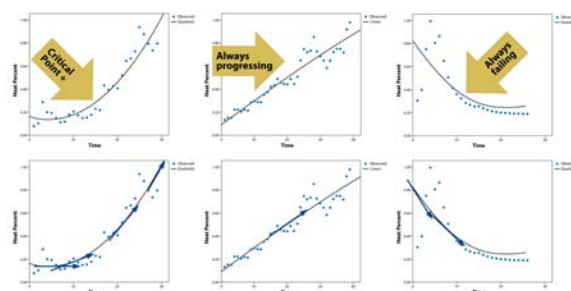


**Learning Dynamics.** It is relatively easy to analyze and interpret the learning dynamics in an individual case study like the one illustrated in Figure 1.3. However, practical application for assessment required an automated method. To this end, CyGaMEs follows quantification of players' learning as percent of progress toward the game goal by curve fitting, individually regressing Timed Report percent of goal attainment for each of the four goal dimensions as the dependent variable on Time using both linear and quadratic equations (see Figures 1.4 and 1.5). Standardized betas are retained for the curve with a significant beta for highest order term.

Learning Dynamic Equations	Interpretation	Quadratic
$f(X) = b_1X^2 + b_2X$	Progress toward goal	
$f'(X) = 2b_1X + b_2$	Rate of progress	
$f''(X) = 2b_1$	Acceleration in progress	Linear
$f(X) = b_2X$	Progress toward goal	
$f'(X) = b_2$	Rate of progress	
$f''(X) = 0$	Acceleration in progress	

**Figure 1.4.** Learning Dynamics Equations. © 2012 Debbie Denise Reese. Used with permission.

This equation represents a player's progress toward that game goal dimension. Then, first and second derivatives are calculated using the retained beta(s). The first derivative is the slope of the progress line. When player data are substituted for X, this equation identifies the player's rate of progress toward the game goal at each Timed Report. This is the rate of learning and knowledge application. Solving the second derivative at each Timed Report indicates any change in the rate of progress, or acceleration progress. Multilevel modeling on *Selene* first and second derivatives controls for nesting of data within players and quantifies the rate of progress (learning and knowledge application) for each of the dimensions. Analyses from 267 players served as the initial norming population, and a second dataset of 90 players replicated those results.



**Figure 1.5.** Examples of learning dynamics curves. The two on the left and right of each row are quadratic, and the center graphs are linear. The arrows in the bottom row illustrate slopes. © 2012 Debbie Denise Reese. Used with permission.

As independent research teams across the world explore instructional game design and embedded assessment, it will be necessary to achieve some standardization across populations, instruments, and methods. How does this new field synthesize conclusions from disparate information? Instructional game assessment will require the invention of standards of measure to permit consistency, precision, technology advancement, and informed decisions.

Learning dynamics are a type of analytics that determines rate of learning progress and acceleration. Learning dynamics can be used to obtain standardized measures to calibrate instructional games and compare:

1. Instructional games for whether or not they effectively cause learning of new knowledge and its application
2. Learners within a game
3. Learners across games

Timed Report measures are measures of knowledge acquisition and application because the measures derive from knowledge domain specification translated into procedural transactions as aligned game world analogs.

## Future Work in A Priori Knowledge Specification

### Alignment Based upon Cognitive Task Analysis

John R. Anderson repeatedly advised that the educational effectiveness of Cognitive Tutors is a function of the accuracy of the task analysis (in concert with the quality of the tutor-based instruction [diagnosis and feedback] and classroom deployment):

The development of an accurate cognitive task analysis has been and continues to be the most time intensive part of developing instruction. This is because we must examine each domain we want to teach anew and cannot carry over knowledge from one domain to another. This is not a unique problem of the tutoring approach. Any attempt to do instruction informed by cognitive models is going to face this investment. As we lamented earlier, there does not seem to be any professional organization to foster development of such cognitive task analyses (Anderson & Schunn, 2000, p. 22).

Unlike Cognitive Tutors, many contemporary approaches to instructional simulations and instructional games lack rigorous alignment (NRC, 2011). Cognitive Tutors, Science ASSISTments, *SimScientists*, and CyGAMES are examples of NSF-supported projects implementing approaches that conduct and apply task analysis. Cyberlearning environments like Cognitive Tutors, simulations, and instructional games afford automated collection of vast amounts of data for quantifying student interactions during authentic performance tasks. These tasks can be engineered as embedded assessments to measure learning and provide formative and summative feedback. However, data are useless unless (1) data contain relevant signal(s), (2) data translate into information, and (3) data reports effectively communicate that information. Data mining methods afford the ability to apply a type of grounded theory approach to quantitative data so as to allow information to emerge from the data. This affordance does not obviate designers' responsibility to conduct rigorous task analysis and ensure alignment among targeted knowledge, the cyberlearning environment, and the embedded assessment measures. Embedded assessment requires a priori alignment of assessment and to-be-learned content through formal knowledge specification. When authentic assessment is embedded within computer-based *instruction*, knowledge specification must align both instruction and the assessment with to-be-learned content. Alignment enables a priori specification of an expert model for comparison to the learner's mental model as evidenced in learner activity (behavior). Computer-based, interactive assessment designs should begin with knowledge specification and apply it to measure learning.

Viable representation of targeted knowledge and skills within instructional environments and their embedded assessments coupled with the methods presented in this paper and the others in this report holds promise for learning analytics that better assess outcomes for enhanced teaching and learning. Today's research lays the foundation for wholesome and effective cyberlearning of the future. Given a student with individual goals, strengths, and weaknesses, instructional systems will one day show students what they already know, what they need to learn, and how they will best learn it. They will scaffold students through knowledge acquisition and show them what they learned, when they learned it, and their options for what to learn next.

---

This material is based upon work supported by the National Science Foundation under Grants Nos. DRL 0733286; DGE 0742503; DRL 0814512; DRL 1008649; DRL 1020264; IIS 8318629; DRL 8470337; PHY 8715890; DRL 8954745; DRL 9253161. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

The author extends appreciation to the CyGAMES staff, partners, and collaborators, recruiters, and players who help make CyGAMES theory a reality, in particular Robert Kosko, Barbara G. Tabachnick, Ben A. Hitt, Ron Magers, Charles A. Wood, Janis Worklan, Cassandra Lightfritz, Victor Hernandez-Gantes, Matt Petrole, Ralph J. Seward, and Steven Nowak. The quality of this work was enhanced through investment in conversation and



reviews by Kenneth R. Koedinger (CMU Cognitive Tutors), Edys Quellmalz (SimScientists), and the NSF DR K-12 CADRE New Measurements Paradigms SIG community.

Correspondence concerning this section of the manuscript should be addressed to Debbie Denise Reese, Center for Educational Technologies, Wheeling Jesuit University, 316 Washington Ave., Wheeling, WV, 26003. E-mail: [debbie@cet.edu](mailto:debbie@cet.edu)



## 2. Item Response Theory and Beyond

- Douglas H. Clements, Julie Sarama—University of Buffalo
- Michael Timms—Australian Council for Educational Research
- Curtis Tatsuoka—Case Western Reserve
- Kikumi Tatsuoka—Columbia University (retired)

Many researcher-developed instruments have used classical test theory and scoring, such as giving each student a score that is the sum of all the correct items. Although seemingly straightforward, there are many problems with this approach. For example, if student A gets a single digit math addition problem like  $6 + 7$  correct, but does not even try a harder multi-digit question like  $397 + 286$ , he receives a total of 1 point toward his total score. In comparison, if student B gets  $6 + 7$  incorrect (say, writing 12 instead of 13 because of a simple memory lapse), but correctly answers  $396 + 287$ , she also receives a total of 1 point toward her total score. Although the raw scores are equal, most people would be confident that, from their responses to these two items, these students are not at the same level of ability in arithmetic.

### Use of Item Response Theory and Other Measurement Methods in Intelligent Learning Environments

#### Item Response Theory

Rasch analyses, and the wider family of “Item Response Theory” (IRT) models<sup>8</sup>, address such problems. These models can be used to estimate the distance between items, as well as between persons, and express them on a common scale. That is, Rasch developed a probabilistic model in which item difficulty based on the proportion of a given sample that responded correctly, and a person measure based on the proportion of items the person completed correctly, are simultaneously estimated. The result is a scale on which both persons and items are mapped onto the theoretical latent trait. A latent trait is an attribute of persons, in our case mathematical competence, that can be inferred from their behaviors (Bond & Fox, 2001; Linacre, 2005; Watson, Callingham, & Kelly, 2007).

Thus, IRT models are *linear* measures on an ordinal scale. Classical scores are not. Of the additional benefits of IRT models, we briefly mention seven (Embretson & Reise, 2000): (1) Data for classical instruments must be complete to be summed. Missing data present serious problems. In IRT, models are robust against missing data, including items not administered. (2) Classical scores provide only ordinal rankings, whereas IRT models provide estimates of each person’s (and item’s) position on a linear scale. Thus, scores are additive and students of different ages can be placed on the same scale meaningfully. Further, in classical scoring, gain scores are typically negatively correlated with initial level, so that children with low pre-test scores artificially show the largest gain. (3) IRT scores’ accuracy and precision can be quantified. (4) For these reasons, IRT scores are better suited for most inferential statistical methods. (5) IRT measures can be equated more easily. (6) IRT models are less sample- and test-dependent than classical models. (7) In IRT models, construct validity is a comprehensive concept that includes content validity, face validity, and concurrent validity (Bond & Fox, 2001; Smith, 2004). Because Rasch modeling is a theory-based approach to developing measures through hypothesis testing (Andrich, 2004; Wilson, 2005), when data fit the Rasch model, they provide evidence of the construct validity of the instrument (Smith, 2004). For example, it can add evidence to the validity of a researcher’s hypothesis of the unidimensional progression of mathematical achievement for a topic or domain.

<sup>8</sup> Some argue Rasch is not a member of the IRT family, as it does not attempt to fit a model to data but requires that the data fit the model. IRT models may have different parameters, such as items having different degrees of discrimination or corrections for guessing.



As examples, several researchers used Item Response Theory (IRT) to develop and evaluate assessments for early mathematics development (de Lemos & Doig, 1999; Doig & de Lemos, 2000; Thomson, Rowe, Underwood, & Peck, 2005). We similarly used the Rasch model to design a measure based on learning trajectories to assess core mathematical abilities of children from age 3 to 9 years (Clements, Sarama, & Liu, 2008). That is, abilities are assessed according to theoretically and empirically based developmental progressions that underlie research-based learning trajectories—that we consider central for diagnostic assessment. The researchers defined mathematical competence as a latent trait, yielding a score that locates children on a common ability scale with a consistent metric. Analysis of the assessment data showed that its reliability ranged from .92 to .94 on the total test scores (Clements, et al., 2008). The final instrument is calibrated to widely used standards and curricula (Clements, Sarama, & Wolfe, 2011).

A study by Timms (2007) used IRT in an intelligent help system to deliver hints to students according to their needs in a tutoring system designed to help middle school students learn to select and use equations when solving physics problems about speed. The tutoring system used IRT to measure the size of the gap between a student's initial ability and the difficulty of tasks that they were about to undertake. This was designed to take advantage of the fact that IRT can be used to scale the ability of the students on the same scale as the difficulty of the items they respond to and, therefore, be used to measure the gap between the difficulty of the item and the student's ability to respond to it. The tutor used this to give hints to students appropriate to the size of their learning gap. The second version of the tutor provided feedback on errors made but gave no hints on how to repair those errors. The third version of the tutor gave neither error feedback nor hints. The results showed a statistically significant difference between the mean learning gains of students using the tutor when compared to a group that received no help but just practiced on the tasks unaided. The difference between the groups represented a moderate standardized effect size of .70.

Although the Rasch model and other IRT models have significant advantages, significant caveats must be considered. Most simply, IRT models are more difficult to compute and to explain. For example, the strength of the Rasch approach is that it requires the data fit a model of fundamental, that is, unidimensional, linear measurement (Bond & Fox, 2001). A weakness is that it assumes that all items have similar discrimination properties, which can be inconsistent with the types and variation of questions desired in an assessment. Thus, the simplest and most complex questions will tend to have greater error in their estimates. Complex interrelationships between different topics that interest some researchers may not fit unidimensional models. There are IRT models that address some of these problems (see multidimensional item response, more complex structural equation models, and other models in Wilson, 2009). Interpretation is also not as straightforward as it may first appear. For example, a gap in the measurement scale might be used to validate separate levels in a learning trajectory. However, these also may alternatively be gaps in the question distribution, with factors other than level of mathematical thinking (as a simple example, complexity of the language) accounting for difficulty.

In summary, even with these caveats in mind, IRT models offer advantages when dealing with complex data sets generated by intelligent learning environments. IRT models yield scores that locate learners on a common ability scale with a consistent, justifiable metric, allowing accurate comparisons, even across ages and meaningful comparison of change scores, even when initial (e.g., pretreatment) scores differ (Wright & Stone, 1979).

## Q-matrix Theory and the Rule Space Method

Despite these advantages, there are several purposes for measuring achievement that IRT models do not fully satisfy. For example, we would like to modify our measure of early mathematics so it can achieve the following.

- Yield information both about each student's overall mathematical progress, but also about their concepts and skills in each of multiple domains or topics of mathematics

- Provide information about specific concepts and skills that constitute the cognitive components of thinking involved in various tasks
- Provide information about general processes involved in successful mathematical thinking across several items and topics
- In so doing, provide profiles of individual children that are more informative than a single score and immediately useful for formative assessment
- Allow a fine-grained evaluation of the theorized developmental progressions and their inter-relationships, thus informing future research, curriculum development, and pedagogical efforts

Both classical test theory and IRT are best employed to obtain an aggregate score across the set of items on a test. However, because different items each measure different concepts, skills, or procedures—or *attributes*—different item response patterns will correspond to different attribute profiles. In fact, there can be large numbers of ways that correct-item combinations can produce the *same* total score or IRT theta. As a result, neither classical nor IRT test theories lend themselves easily to providing cognitive diagnoses of a learners' strengths and weaknesses.

To do so, we are presently using Q-matrix Theory and the associated Rule Space Method (RSM), which are based on the notion that successful performance on tasks, including those in mathematics, is dependent on multidimensional abilities (K. K. Tatsuoka, 2009). This is conceptually similar to conducting multidimensional measurement. Q-matrix Theory involves identifying cognitive attributes that play important roles in performance on individual test items. This is done through cognitive analysis of each item, and identifying what is required to perform well. With a specified link between items and attributes, as denoted by what is known as a Q-matrix, it becomes possible to assess learners' strengths and weaknesses on the identified attributes from observed item responses. RSM is then used iteratively to evaluate and refine the Q-matrix, and in so doing, simultaneously provide a direct test of the theoretical components on which it was based. Both empirical support for the hypothesized learning trajectories and any refinements of them will make significant contributions not only to research in educational psychology and math education, but also to educational standards, curriculum development, pedagogy, and professional development (Clements & Sarama, 2009; Sarama & Clements, 2009; K. K. Tatsuoka, 2009).

## Poset Models, RSM, and Computer Adaptive Testing

The fundamental idea of IRT's latent trait theory is that an individual's behavior can be accounted for by defining certain human characteristics called *traits*, and hence, we estimate quantitatively the individuals' standing on a trait, such as mathematical competence. However, no further diagnostic information is available. In contrast to IRT, cognitively realistic and worthwhile psychometric models and statistical methods should handle multidimensional latent abilities characterized by many unobservable attributes. Q-Matrix Theory and the RSM use IRT analyses but go beyond IRT models to make inferences about several numbers of attributes from observable item responses.

The underlying profiles generated from a Q-matrix form classification states that are partially ordered (C. Tatsuoka, 1996, 2002). Therefore, one can use a finite partially ordered set, or *poset*, approach to classify students efficiently (C. Tatsuoka & Ferguson, 2003). Compared to IRT models, poset methods are discrete, which leads to more efficient and decisive classification, particularly in adaptive testing. We are using RSM and poset methods in developing the Q-matrix, and then final poset models will be fit and validated for computer adaptive testing, saving up to half the time of test administration.



## Future Work

A major advantage of IRT models is that they can locate persons and learning tasks on the same scale, enabling direct comparisons that allow inferences to be made about how easy or difficult a learner might find a task. Future work is needed to explore how this property might be used in intelligent learning environments to recommend the next instructional step for a student or to deliver assistance that is pitched at the correct degree of help the learner needs. IRT models also may have a place in validation of other methods, such as checking if how a Bayes Net categorizes learners is supported by evidence from a multidimensional IRT modeling of the same data. Similarly, researchers could investigate how Q-matrix and poset models might also be used in this way. IRT might also provide quantified estimates of difficulty of learning tasks that can then be used as evidence of how to set the prior conditional probabilities of a Bayes Net that is to be used in dynamic measurement of students as they work through an intelligent learning environment.

---

This material is based upon work supported by the National Science Foundation under Grant No. DRL 1019925. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

### 3. Machine Learning Methods

- *Michael Timms—Australian Council for Educational Research*
- *James Lester, Kristy Elizabeth Boyer, Eric Wiebe—North Carolina State University*

Machine learning is one of the key elements of artificial intelligence (AI) and describes the function of how a computer can adapt to new circumstances by detecting patterns and extrapolating from them (Russell & Norvig, 2010). As assessment of learning moves inexorably to become embedded in more complex tasks that are delivered by computer, the need to adopt techniques from the field of AI increases. Machine learning is a broad area that uses a variety of methods to derive behaviors from empirical data, such as from sensors or, in the case of educational uses, from databases of actions taken by students in learning environments. Approaches take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. The relations between observed variables are captured in the data that they yield. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too large to be covered by the set of observed examples (training data). Hence, the challenge is to generalize from the given examples so as to be able to produce a useful output when processing new cases. Obviously, the quality of the generalizations depends upon the quality of the data gathered in the first place. Methods of learning patterns from data also vary from more exploratory (bottom up) procedures to more theoretically driven approaches (top down), and the outcomes will vary based on methods selected. With some phenomena, it is important that input data (e.g., students' inquiry skills data) are selected and aggregated in a theoretically grounded way, and so it is important to make informed decisions about what data to include and at what level of granularity.

The field of machine learning is broad, and this paper is not intended as a survey of all available techniques. Rather, this paper describes how three particular methods—Bayes Nets, Dynamic Bayes Nets, and Hidden Markov Models—are being used for measuring learning in electronic environments developed in the projects sponsored by the National Science Foundation's REESE and DR K-12 programs.

#### Methods Used in Intelligent Learning Environments

##### Bayes Nets

Bayesian Networks, or Bayes Nets (BNs) for short, have a long history of use in electronic environments, dating back to Judea Pearl, who first developed techniques to build BNs (Pearl, 1988). BNs are used in a wide variety of ways in fields as diverse as astronomy and speech recognition. In this short section, we describe how they are used in a current sample of e-learning and e-assessment projects to measure complex learning and to discuss their advantages and limitations. BNs, in this setting, form a class of Diagnostic Classification Models (Rupp, Templin, & Henson, 2010) that have been widely used in intelligent tutoring systems to predict student behavior and make tutoring decisions (Woolf, 2009), and their use in systems for assessment is growing. Martin and VanLehn (1995) and Mislevy and Gitomer (1996) studied the applications of BNs for student assessment. Mislevy has continued this work with Behrens in the NetPass program, which assesses examinees' ability to design and troubleshoot computer networks (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008). Conati, Gertner, and VanLehn (2002) applied BNs to both assessing students' competence and recognizing students' intentions.

Within the CADRE projects that have been funded by the National Science Foundation, BNs are currently being used in various ways. Researchers at WestEd are using a BN system (among other methods) in the



*SimScientists* simulation-based science assessment (<http://www.simscientists.org>). Crystal Island, an intelligent game-based learning environment developed by the IntelliMedia group at North Carolina State University (<http://www.intellimedia.ncsu.edu/projects.html>), uses Dynamic Bayes Nets to model learning and provide hints to eighth-grade science students. The ASSISTment system developed by researchers at Worcester Polytechnic Institute uses BNs to estimate probability that students possess particular math skills based upon their responses to multiple-choice questions from Massachusetts' state test and other scaffolding questions. The JavaTutor project, also part of the IntelliMedia group at NCSU, is using a related method to BNs that is attempting to learn Hidden Markov Models (HMMs) to discover the structure of task-oriented tutorial dialogue in order to develop a dialogue-based tutor to teach Java programming.

A BN represents a set of random variables and their conditional independencies in a probabilistic graphical model shown via a directed acyclic graph. In the BN, nodes represent random variables, and the edges (links between the nodes) encode the conditional dependencies between the variables. Across a series of nodes and edges, a joint probability distribution can be specified over a set of discrete random variables. Figure 3.1 shows an example of a fragment of a BN used in the scoring of the ecosystems benchmark assessments in *SimScientists*. It shows how observation nodes in the network representing data gathered from student actions in the assessment (the lower two rows) provide information to assess the upper level “hidden” (not directly observable) random variables denoting content knowledge and science inquiry skills represented in the upper two rows.

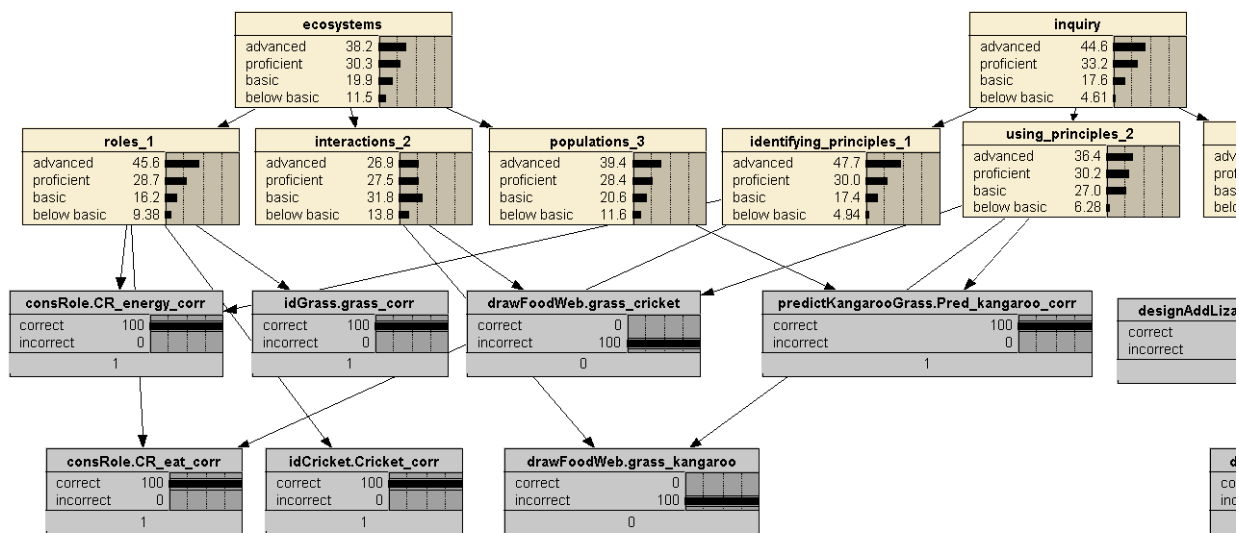


Figure 3.1. Fragment of a Bayes Net from the *SimScientists* Ecosystems Benchmark Assessment.

Values for the edges are encoded in conditional probability tables, but not visible in this view. Data are gathered from student interactions with the simulation or game and passed to the BN where estimation algorithms are then applied using software such as Netica to perform inference to produce estimates of probability that students possess the knowledge or skill represented via the hidden nodes.

### Advantages of Bayes Nets

BNs have several advantages for measurement in complex electronic learning and assessment materials. First, BNs can be used where dynamic assessment is needed to decide, for example, what hints a student might need or which is the optimal instructional step to give next. Because they model how the evidence gathered from a student's actions and responses during the learning or assessment process is related to the variables that represent the knowledge and skills being monitored, a single piece of evidence can be used to update the estimate of the degree to which a student has demonstrated the particular knowledge or skill.

This ability to make estimates based on limited data enables decisions to be made about what to do next for that student, even though our confidence in the estimate may be lower because it is based only on a few observations. For giving hints and making decisions about the next instructional activity, this is very useful because such actions can tolerate some lack of reliability in the estimation. A student's progress is unlikely to be marred by getting a hint that she might not have needed at that point, for example.

A second advantage is that BNs are particularly useful in measurement situations where there is a lot of item dependence because items are all related to particular tasks, for example, in simulation-based assessment tasks where a student makes a series of responses or actions based on one scenario. BN models offer a large degree of flexibility for modeling dependence among observables (item outcome variables) from the same task, which may be dependent. For example, if two skills are regarded as independent, but an observable event gives evidence that can be used to update beliefs about the student possessing each skill, then in the BN this relationship can be represented by connecting the nodes representing the two skills. The graphical display format of the BN can aid in model specification. Another way of modeling dependency, such as that created by the fact that a series of observations of different skills were made in the same scenario, is to create a node in the BN that represents a "context" variable.

A third advantage of BNs is their ability to model dichotomous and continuous latent variables (including polytomously scored items) data as well as dichotomous and polytomous outcome variables. Particularly useful is that BNs can be used to model variables where there is not a linear progression. For example, time spent on a task might be modeled in seconds and there might be an optimum amount of time to spend on it, where spending too much time on it represents a student struggling and spending too little time means that the student has not fully considered the complexity of the task. In traditional educational measurement, an increase in the value of a variable is typically taken to represent increased performance, and that approach does not model this kind of pattern of performance as easily as a BN can. What is more is that a BN can handle observations from a range of dichotomous, polytomous, and non-linear sources simultaneously.

A fourth advantage of BNs is that they can model the complexity of relationships among the latent variables. This is different from the complexity arising from dependencies among items that is discussed above. For example, a BN can be constructed to both handle compensatory and non-compensatory models of the latent variables being measured. Compensatory models are those in which a student's lack of mastery of one skill can be made up for by their mastery of a different skill. This can be modeled by connecting the two skill nodes. In contrast, a non-compensatory model is one where the lack of mastery on a skill cannot be made up for by a mastery of a different skill. In this case, no link would exist between the nodes for the two non-compensatory skills. A single BN can even model a mix of compensatory and non-compensatory relationships.

The fifth advantage of BNs is that they can be used with a wide variety of item types, including conventional assessment items like multiple-choice questions. For example, the ASSISTments project sets conditional probability values ad hoc to represent the fact that a student who does not possess the skill being measured might just guess the correct answer and, conversely, that a student who does have the skill might make a mistake and pick an incorrect answer choice (Pardos, Heffernan, Anderson, & Heffernan, 2010).

### *How Data Is Gathered and Fed to the Bayes Net*

Data gathering for processing in the BN happens in a variety of ways. One way is to gather data as the learner interacts with objects in the learning environment, such as drawing an arrow linking one organism on a food web to another. Another source of data is from making selections, such as answering a multiple-choice option or checking boxes of a list of things to be included in a task. Other possible data sources are



whether a student reaches a particular screen or hotspot on a screen, and the response of a learner to an open-ended question where a text or number entry is made that can be evaluated for correctness. At its simplest level, a binary 0/1 response is captured, but the network can also handle input of polytomous items where a scale might be 0/1/2/3 or a continuous variable like time taken in seconds.

Updating of the BN can happen dynamically so that each new piece of evidence is used to propagate through the network and update the estimates of the variables being measured. This has advantages of maintaining the most up-to-date estimate of the student's ability or level of mastery, which can be useful if frequent tutoring intervention needs to take place and the student's ability is changing over time. The disadvantage is that there is some processing time to the calculation, especially when the BN is large. Also, one additional piece of evidence on its own may not make much difference in the estimates of ability or mastery. So, it is common to send the data to the BN in batches, which will more significantly affect the measured outcomes and will speed up the transaction time. Similarly, in some cases, the data are used to update just a fragment (subpart) of the larger BN that is relevant to the particular task in which the student is currently engaged.

BNs can be used to gather information on the learner's performance over time as well as provide immediate feedback. As the student works through series of tasks over time, data from each can be used to update the estimates produced by the BN to classify student performances. For example, individual tasks might only update the student classification for the relevant fragment of a larger BN, but when several tasks are completed, the fragments together provide sufficient information to provide more accurate estimates across the whole BN.

### **Dynamic Bayes Nets (DBNs)**

A related use of BNs to model the learner across time is dynamic Bayes Nets (DBNs), which are a type of BN that follow common structural patterns in order to incorporate time. BNs typically have fixed structures and are well-suited for modeling static processes. During modeling, no new nodes or links are traditionally added to the network. In contrast, DBNs are well suited for modeling temporal processes, where a random variable's value is likely to change over time. DBNs append structurally equivalent "slices" to growing networks that model processes over time.

In BNs, whenever a random variable is observed to have a new value, the associated node in the network is "clamped" (set) to that value and information about the variable's previous state is lost. In DBNs, each random variable has a corresponding node for every moment in time. Nodes are still clamped as observations are made. However, previous model states can indirectly influence future model states in DBNs. Variable history is not "overwritten" in the same way that it is in BNs. When a DBN is "unrolled" (all slices are explicitly represented), a DBN is just a large BN. Inference and interpretation can be performed in the same way, although practical requirements may require the use of approximate inference techniques for DBNs. Furthermore, DBNs are a generalization of other temporal graphical models, such as Hidden Markov Models, which are discussed later in this paper.

One of the key challenges in using DBNs for knowledge assessment is identifying knowledge components (see the earlier paper on Knowledge Specification in this document) and the relationships among them. Structure learning is notoriously hard in BNs, so DBNs for knowledge assessment are generally hand-authored. One can specify a set of assumptions and procedures to systematically choose nodes and directed links. However, the inherently unobservable nature of student knowledge makes validation of these models difficult. As for identifying conditional probabilities to parameterize the models, manual assignment often requires arbitrary decision-making, but machine learning the parameters is very slow for all but the simplest models.



An example of the use of DBNs is in Crystal Island (Rowe & Lester, 2010), in which hidden nodes represent knowledge components from four relevant categories: narrative knowledge, content (curriculum) knowledge, strategy knowledge, and scenario solution knowledge. Observed nodes correspond to concrete actions taken by the student in the virtual environment, such as speaking with virtual characters or performing particular lab tests. The idea is to treat concrete in-game actions as evidence of student knowledge through the framework of DBNs. Each time the user takes a substantive action, a new slice is appended to the network and the evidence is propagated through the model. Then the model's final knowledge estimates are used to predict students' performance on a curriculum post-test.

### Future Work with Bayes Nets

One of the challenges with BNs is how to validate their output. In an “expert centric” approach, a BN is constructed based on existing content knowledge about the domain by experts. At the outset, the nodes and edges are established by the designer of the learning environment and, typically, the prior probabilities are set by expert judgment. Because the BNs can quickly become quite complex with many nodes and relationships among them, it can be difficult to easily establish whether the BN is correctly classifying learning performances as was intended. To do so, pilot-test data needs to be gathered from real students to generate a dataset that can be used for validation. It is also necessary to have an independent measure or estimate of the performance of the students on the variable of interest, or something very close. This might come from an external measure, such as their performance on a state test of the same subject or from a teacher's judgment of the students' abilities or accomplishments in the topic. Then correlational studies of the student performance, as classified by the BN compared to the external measure, can be conducted to see if they are classifying in similar (but not exactly the same) ways. Another way to validate the classifications made by the BN is to use other measurement modeling methods to analyze the dataset. For example, the *SimScientists* program has used multidimensional Item Response Theory (see the previous paper for more information on IRT) to find the best-fitting model to categorize the student performance as represented in the dataset. Then expert judgment will be used to determine cut points between the four levels of performance. The best fitting model was then used to retrain the BN, a process in which the students' category of performance is treated as known (from the IRT analysis) and the BN then “learns” from the scored dataset the patterns of responses of students in each category. Another method for validation involves having experts look at the student outcomes from the IRT model and their actual performance on different parts of the task and then to adjust the prior probabilities based on the patterns they observe. This is a more time consuming method than the automated retraining.

A different approach to constructing the BN is called *data centric*, and begins with pilot testing the e-learning environment with students to obtain real data from which the BN can be developed. Then, Educational Data Mining methods (see the next paper in this report for more information on Educational Data Mining) can be used to derive the structure of nodes and the conditional probabilities of the edges in the BN. Even learning from data has its constraints, though, and researchers may end up making trial-and-error adjustments to tweak conditional probability tables, as happened in the case of the DBN that underlies Crystal Island, in which students are likely to conduct multiple tests, read books multiple times, speak with characters multiple times, etc., in an expansive and open-ended learning environment that makes it impossible to enumerate all problem-solving paths through the game.

### Hidden Markov Models

Hidden Markov Models (HMMs) constitute a stochastic approach to characterizing the observed signals emitted from a source (Rabiner, 1989). The premise of HMMs is that some aspect of the signal source is hidden (i.e., not directly observable) and that the values of this hidden variable (i.e., the hidden states) are important for modeling the system as a whole. The model is said to be “in” one of the  $N$  hidden states at



each step in the observed sequence. Each hidden state is characterized by a probability distribution over the observed symbols called the *emission probability distribution*, while the transitions among hidden states are governed by the *transition probability distribution*. When training an HMM to model a particular phenomenon, the goal is to select the model that maximizes the probability of the observed input.

### *Advantages of Hidden Markov Models*

HMMs have been applied successfully to a range of tasks in intelligent tutoring systems research. These include modeling student activity patterns (Beal, Mitra, & Cohen, 2007; Jeong et al., 2008), characterizing the success of collaborative peer dialogues (Soller & Stevens, 2007), and learning human-interpretable models of tutoring strategies, or modes, in tutorial dialogue (Boyer et al., 2009). In the context of tutorial dialogue, we can formulate *observation symbols*, constituents of the input sequences from which the model is learned, by annotating dialogue. An example annotation scheme, given from a corpus of tutorial dialogue for introductory computing (the JavaTutor '08 corpus) is displayed in Table 3.1. Each observation symbol is said to be “generated” by a hidden state according to that hidden state’s emission probability distribution. Hidden states can be interpreted as tutoring modes and given names pursuant to that interpretation, such as “Student Acting on Tutor Help,” which is characterized by a probability distribution of dialogue acts and task actions that would be observed when the tutorial dialogue is in that hidden state (Figure 3.1). These may correspond to a notion of tutoring strategies, or “modes” (Cade, Copeland, Person, & D’Mello, 2008).

An underlying premise of applying HMMs to tutorial dialogue is that natural language dialogue is influenced by a layer of unobservable stochastic structure. This structure is likely comprised of cognitive, affective, and social influences, among others. In other words, there are latent constructs that influence the performance on tasks in a probabilistic fashion. HMMs provide a framework for explicitly modeling unobservable structure within a layer of hidden states. This modeling framework for extracting tutoring modes and analyzing their differential effectiveness has direct applications in authoring data-driven tutorial dialogue system behavior and in research regarding the effectiveness of human and machine tutors alike.

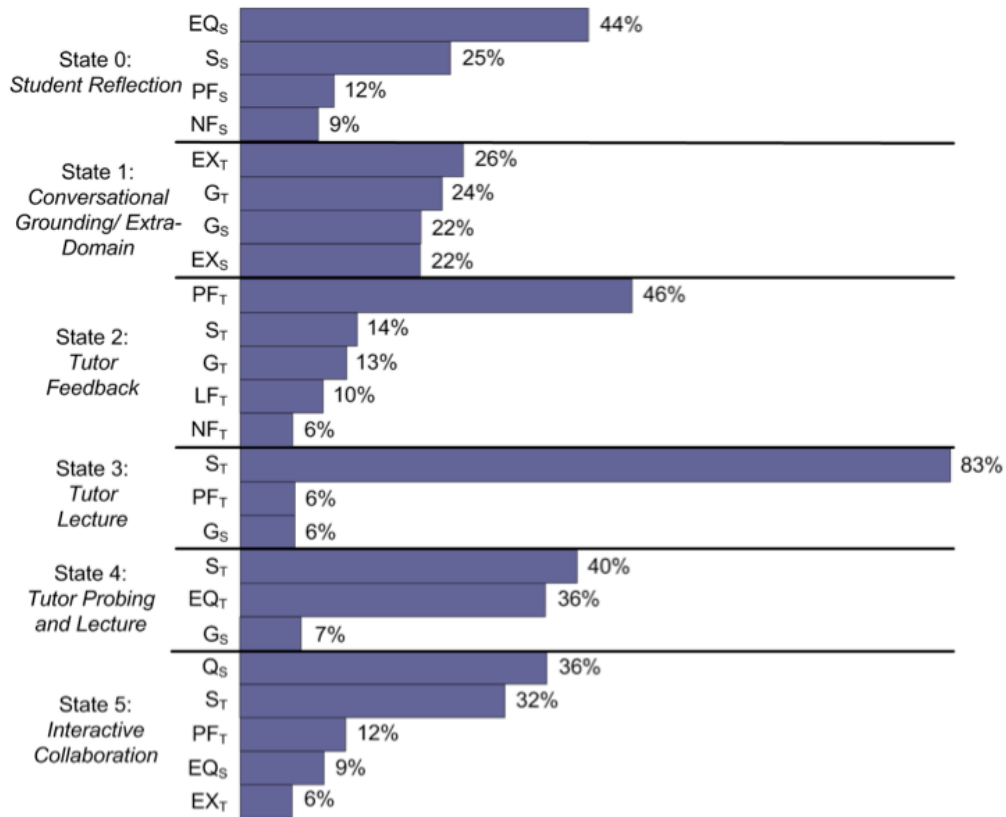
**Table 3.1.** Example dialogue act annotation scheme.

Act	Description	Tutor and Student Example Utterances	Relative Frequency Across Corpus (%)	
			Student	Tutor
Question (Q)	Questions about goals to pursue, domain concepts, etc.	"Where should we start?" "How do I declare an array?"	5.72	0.53
Evaluative Question (EQ)	Questions that explicitly inquire about student knowledge state or correctness of problem-solving action.	"Do you know how to declare an array?" "Is that right?"	8.55	6.15
Statement (S)	Declarative assertion.	"You need a closing bracket there." "I am looking for where this method is declared.."	4.34	40.85
Grounding (G)	Conversational grounding.	"Alright." or "Okay." "Thanks." or "Hello."	5.12	3.72
Extra Domain (EX)	A statement not related to the computer science discussion.	"The problem description is on your desk." "Can I use my book?"	2.75	3.58
Positive Feedback (PF)	Unmitigated positive feedback regarding problem solving action or student knowledge state.	"Yes, I know how to declare an array." "That is right."	2.38	10.63
Lukewarm Feedback (LF)	Partly positive, partly negative feedback regarding student problem solving action or student knowledge state.	"Sort of." "You're close." or "Well, almost."	0.66	2.01
Negative Feedback (NF)	Negative feedback regarding student problem solving action or student knowledge state.	"No." "Actually, that won't work."	1.89	1.11

Learning an HMM involves training its emission, transition, and initial probability distributions to maximize the probability of seeing the observed data given the model. Model training is an iterative process that terminates when the model parameters have converged or when a pre-specified number of iterations has been completed. The fit of a model is measured with log-likelihood, which has a monotonic relationship with likelihood and is less susceptible to numerical underflow. The iterative training process operates over a particular number  $N$  of hidden states. The training approach in the Crystal Island project uses the Baum-Welch iterative training algorithm (Rabiner, 1989) and incorporates a meta-level learning procedure that varies the number of hidden states from 2 to 20 and selects the model size that achieves the best average log-likelihood fit in ten-fold cross-validation.



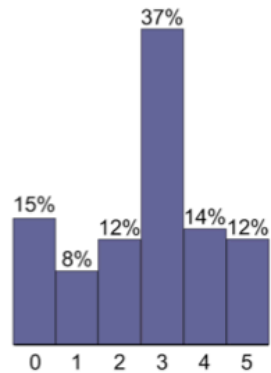
(A) Dialogue Act Emission Probability Distributions by Dialogue Mode\*



\* Emission probabilities less than 5% are not depicted.

State $t$	State $t+1$					
	0	1	2	3	4	5
0	0.083	0.017	<b>0.712</b>	0.013	0.064	<b>0.107</b>
1	0.054	<b>0.746</b>	<.001	<.001	0.075	<b>0.126</b>
2	0.096	0.016	0.015	<b>0.482</b>	<b>0.238</b>	<b>0.15</b>
3	0.036	0.011	0.004	<b>0.437</b>	<b>0.216</b>	<b>0.239</b>
4	<b>0.863</b>	0.011	0.022	0.006	0.064	0.031
5	<.001	<.001	0.016	<b>0.847</b>	0.088	0.046

(B) Dialogue Mode Transition Matrix

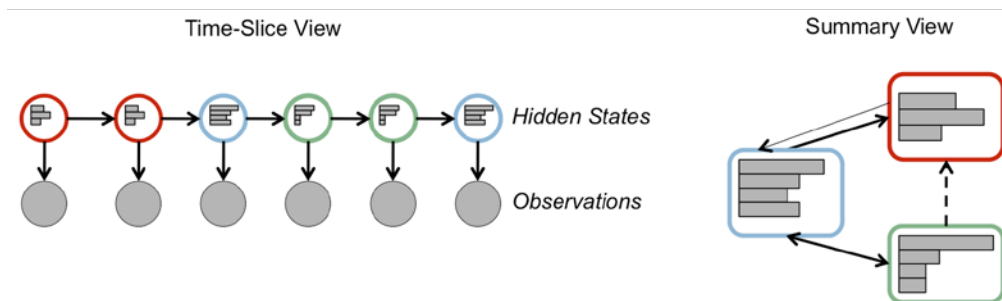


(C) Relative Frequencies of Dialogue Modes

Figure 3.2. Example of (A) emission and (B) transition probabilities for a tutorial dialogue HMM. Relative frequencies of the hidden states, as fit across tutoring sessions, are shown in (C).

HMMs can be graphically depicted in several different ways. Each depiction explicitly shows some components of the model while omitting others for simplicity. Perhaps the most common depiction is a time-slice view (Figure 3.2, left side), in which each time step is displayed along with its associated hidden state and observation. Missing from such a depiction is the probability of transitioning from one hidden state to the next, because in a time slice view each transition did occur. In this paper, the HMMs that have the best fit to the data are also displayed in a summary, or Bayesian, view as shown in Figure

3.3, right side. This view depicts the hidden states as nodes within a graph, and the emission probability distribution of each hidden state as a histogram within the node. Arrows indicate transition probability distributions between hidden states, with heavier arrow weight indicating higher probability.



**Figure 3.3.** Time-slice and summary views of example HMM

## Using Hidden Markov Models for Tutorial Dialogue

A key issue in intelligent tutoring systems research is identifying effective tutoring strategies to support student learning. It has been long recognized that human tutoring offers a valuable model of effective tutorial strategies. Tutorial interactions in both human-human and human-computer learning environments are often studied by collecting a record of the student-tutor interaction (e.g., dialogue transcripts, student action traces). This record constitutes the observable behavior that results from the tutorial interaction, but in many instances the variable of interest (e.g., tutorial dialogue strategy, student engagement) cannot be directly observed in the data. HMMs are well suited to such cases because they represent the unobservable variable (or “hidden states”) as probabilistic distributions over the observed values. The primary benefit of using HMMs to discover these modes lies with the feasibility offered by a bottom-up, data-driven approach in which the theoretical framework is used to devise a set of dialogue act tags that are applied at a low level—usually within a window of approximately one dialogue turn—and then machine learning techniques aggregate the individual dialogue act tags into higher level modes. This methodology addresses an important limitation of the contrasting top-down approach, namely, that sophisticated tutoring strategies rarely occur with novice tutors; for example, recent findings suggest that some widely recognized strategies (e.g., Model-Scaffold-Fade) may not occur fully intact even with highly skilled human tutors (Cade et al., 2008). Identifying dialogue modes with HMMs circumvents the need to manually “design” tutorial strategies and offers an opportunity to automatically discover which strategies are in fact used in practice.

A rich history of tutorial dialogue research has identified some components of these strategies including adaptive cognitive scaffolding, motivational support, and collaborative dialogue patterns that support learning through tutoring (M. T. H. Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Fox, 1993; Graesser, Person, & Magliano, 1995; Lepper, Woolverton, Mumme, & Gurtner, 1993). As the field has grown in its understanding of fundamentally effective tutoring phenomena, it has also become clear that the complexities of choosing the most effective contextualized tutoring strategies are not well understood. An important research direction is to use dialogue corpora, collected from human-human or human-computer tutorial dialogue, to create models that can assess the differential effectiveness of tutoring strategies at a fine-grained level (M. Chi, Jordan, VanLehn, & Litman, 2009; Ohlsson et al., 2007). This research direction is particularly timely given the increasing attention that is being given to machine learning techniques to address tasks such as automatically authoring intelligent tutoring system behavior (Barnes & Stamper, 2010) along with creating data-driven natural language dialogue management systems for tutoring (e.g., Tetreault & Litman, 2008) and other dialogue application areas (Bangalore, Di Fabrizio, & Stent, 2008). Because machine-learned models constitute stochastic representations of



tutoring expertise, they can be used not only in a generative context to author tutoring system behavior, but also as descriptive models whose associations with student outcomes give insight into the effectiveness of tutoring. This is the goal of using HMMs to model tutorial dialogue.

There is growing evidence that meaningful tutorial dialogue patterns can be automatically extracted from corpora of human tutoring using machine learning techniques, of which HMMs are an example (e.g., M. Chi et al., 2011; Forbes-Riley & Litman, 2009; Fossati, Di Eugenio, Ohlsson, Brown, & Chen, 2010; Kersey, Di Eugenio, Jordan, & Katz, 2009; Ohlsson et al., 2007; Tetreault & Litman, 2008). The meaningfulness of these models can be assessed in several different ways, including whether their components are correlated with student outcomes in an existing dataset or whether implementing them can improve the effectiveness of a tutoring system.

### **Future Work with Hidden Markov Models**

Future work in applying HMMs includes taking into account a wider variety of features for modeling; for example, expanding the input sequences from dialogue acts alone to include surface-level utterance content. In addition, knowledge of the task state within tutoring can be used to segment the dialogue in meaningful ways to further refine the structure of the HMM. It is also possible that dialogue tagging at different granularities could reveal varying and useful models. Moreover, the HMM approach can be used to compare tutorial strategies for effectiveness by correlating hidden state usage with outcomes of interest, and by training models separately for students in different groups, as in Jeong, Gupta, Roscoe, Wagster, Biswas, & Schwartz (2008). Using clustering, as in Soller & Stevens (2007), and a finer grained knowledge model could also reveal more detailed tutoring strategies. Finally, combining a bottom-up approach with a top-down approach offers promising synergies. By combining these methodologies, it is hoped that a data-driven understanding of the impact of tutoring strategies on student learning and affect can be formed. This understanding will inform the development of next-generation advanced learning technologies.

---

This material is based upon work supported by the National Science Foundation under Grant Nos. DRL 0733286; DGE 0742503; IIS 0812291; DRL 0822200; DRL 1008649; DRL 1020264; DRL 1138497. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

## 4. Applying Educational Data Mining in E-Learning Environments

- *Diane Jass Ketelhut—University of Maryland-College Park*
- *Alexander Yates, Avirup Sil—Temple University*
- *Michael Timms—Australian Council for Educational Research*

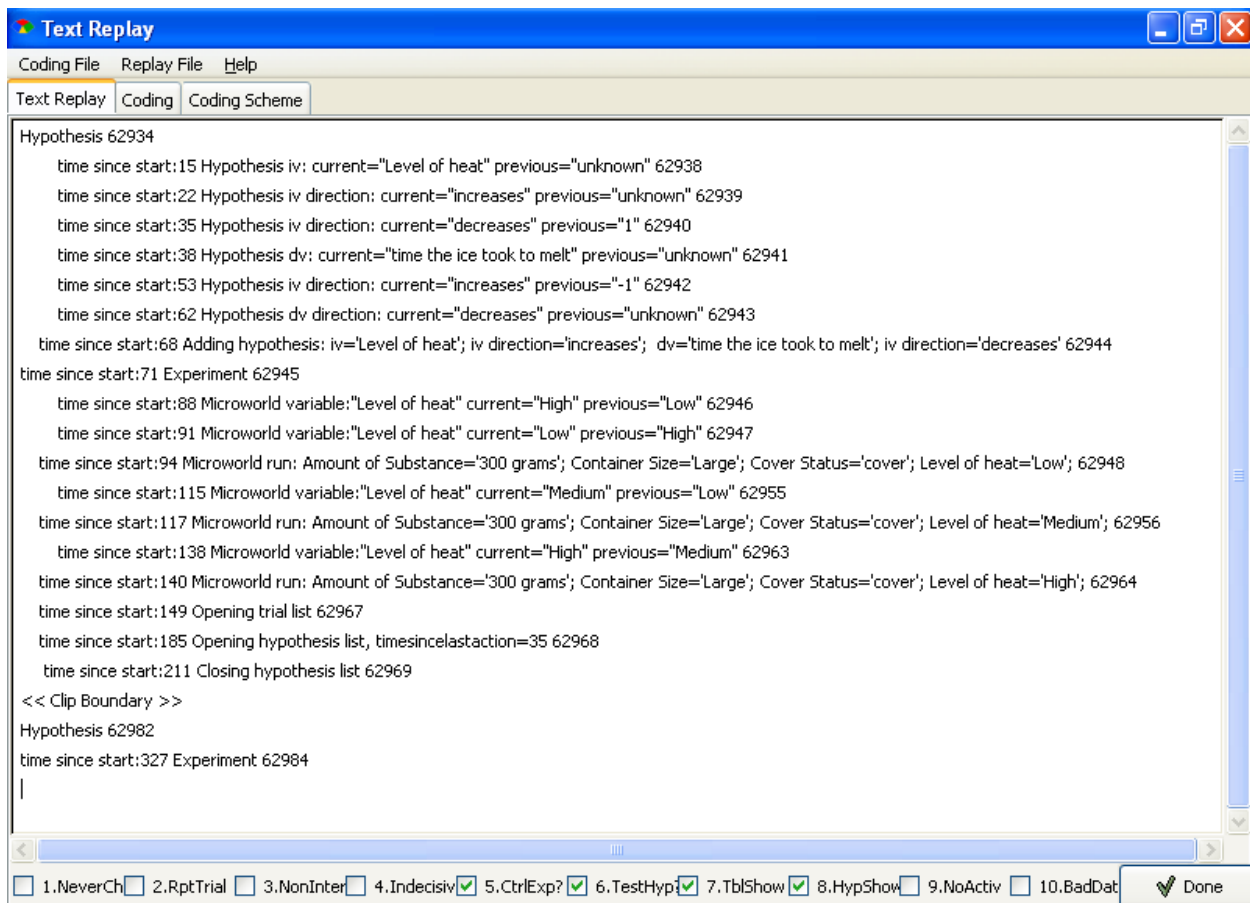
This paper discusses the application of tools and techniques from educational data mining (EDM), an emerging field that spans the disciplines of computer science, statistics, data mining, and educational measurement and focuses primarily on modeling and uncovering patterns in large datasets (Romero, Ventura, Pechenizkiy, & Baker, 2011; Hand, Blunt, Kelly, & Adams, 2000). EDM covers a wide range of methods, and this paper focuses on some examples drawn from a selection of projects funded by the National Science Foundation and connected through the CADRE network.

As the number and complexity of e-learning environments increases to include such rich environments as simulations and games, so does the availability and depth of data on student behaviors, actions, and responses. Understanding what these data can tell us about student understanding is a field in its infancy. This is driving a need for those interested in detecting patterns of learning from the datasets to seek out methodologies from a variety of sources and apply them in new ways (Quellmalz, Timms, & Schneider, 2009).

There are multiple approaches one might take to uncover these patterns. For example, the research team at Worcester Polytechnic Institute has used data mining techniques to improve the use of Bayesian Knowledge Tracing (KT) models (Pardos & Heffernan, 2010). The KT models are commonly used in cognitive tutoring in order to determine student knowledge based on four parameters: learn rate, prior knowledge, guess, and slip. In the past, the Expectation Maximization (EM) algorithm has been used to derive values for those parameters, but previous research showed that with four free parameters the standard KT model is prone to converging to erroneous degenerate states depending on the initial values of these four parameters. This work reports on how data were simulated from a model with known parameter values and a grid search run over the parameter initialization space of KT to map out which initial values lead to erroneous learned parameters. From that analysis, researchers developed an individualization model that has unique properties, allowing it to avoid the local maxima problem.

In other projects at Worcester Polytechnic Institute, the Science ASSISTments research group (<http://www.scienceassistments.org>; Sao Pedro, Baker, Gobert, Montalvo, & Nakama, in press; Gobert, Sao Pedro, Baker, Toto, & Montalvo, accepted) has also used educational data mining to develop detectors to assess students' inquiry skills in real time. Specifically, student inquiry skills were discovered from data in a bottom-up fashion. Models were trained using labels generated through a new method of manually hand-coding log files, "text replay tagging." Figure 4.1 shows an example of a clip from a log file that has been hand-coded. In this method, researchers code segments of log data, machine learning is used to produce a "detector," data and codes are input to machine learned models, the output is a detector for each coding category (which is a complex set of rules or equations), detectors are validated on new datasets, and then they are used to auto-score students' data. This approach led to detectors that can automatically and accurately identify these inquiry skills under student-level cross-validation, providing measures of validity and reliability. The detectors developed can assess whether students are using the control for variables strategy 85% of the time, and whether students are testing their hypotheses 86% of the time. In addition, the detectors can accurately identify these inquiry skills for new students and can capture students' inquiry performances and auto-score them in a manner that handles their complexity.





**Figure 4.1.** An example slip labelled by a human coder. This clip was tagged as involving designing controlled experiments and testing stated hypotheses in addition to other behaviors.

For example, in the behavior models developed, detectors are able to differentiate whether a student knows the control for variables strategy even when the student chooses not to conduct sequential trials. Their approach assesses each inquiry skill over multiple trials, in the context of rich science inquiry microworlds, thereby providing a solution to three previously acknowledged problems: (1) the amount of data required to establish reliable measurement of inquiry skills (Shavelson, Wiley, & Ruiz-Primo, 1999), (2) the capacity to assess inquiry skills in the context in which they are developing (Mislevy et al., 2003), and (3) issues about the validity of measuring skills that are aggregated over multiple task performances (Williamson, Mislevy, & Bejar, 2006).

A third example, Situated Assessment using Virtual Assessment of Science Content and Inquiry (SAVE Science), derives from a multi-university research team at the University of Maryland, Temple University, and Arizona State University. This team has developed a game environment, Scientopolis, to assess middle school children's understanding of both science content and process that has been previously taught in the classroom outside of the environment. Scientopolis has four assessment modules based on unique content and process skills. Children respond to a problem-based narrative in each by exploring the environment, interacting with people and objects in the environment, measuring and collecting possible clues, and using their understanding of the topic and process to draw inferences about the problem (Ketelhut, Nelson, Schifter, & Kim, 2010). SAVE Science allows students to solve the problem in multiple ways; many of these are equally correct while others uncover misconceptions held by the student. It is the goal of this project to develop an understanding of what students have learned about the topic from how they attack the posed problem. To do this, all student interactions are stored in a



database with a time stamp. Teachers are provided with individual student data through the teacher dashboard nearly immediately to use for curriculum planning and student assessment purposes as they choose.

Data mining techniques fall into two categories (Hitt, 2012). While the names for these two categories appear to vary, in essence one category attempts to create predictive models for a collection of dependent variables within a dataset based on a theoretical foundation. The second is based on fewer assumptions and looks for patterns and clusters within the data. Given the goals of the SAVE Science project, both aspects of data mining were employed. In 2010–11, the SAVE Science dataset recorded 59,374 interactions between 562 students (from 8 different classrooms) and the 1,052 items in the three virtual environments that were implemented that year. In addition, there were 16,281 measurements recorded to the clipboard, a virtual notebook. Clearly, the task of discovering patterns is complicated by the size of the dataset as well as by the fact that there are not clear right and wrong pathways to solving the problems.

Four steps are being conducted to understand this dataset: principal components analysis (PCA), correlations, clustering, and visualization. Each will be described below with an example of what we have learned from it.

## Principal Components Analysis

PCA is a standard data mining technique used for a variety of purposes, including dimensionality reduction, feature selection, and data visualization (Romero et al, 2011). It was used in SAVE Science to identify the most prominent directions (or dimensions) of variation in the dataset, which will help identify features that are highly correlated with one another, and subsets of the feature set that are mostly independent from one another.

Given a set of points with zero mean, the first principal component identifies a linear regression of the points—the vector that minimizes the sum of the squared distance between each point and the line given by the vector. The second principal component identifies another linear regression, but this time a regression of the points after the first principal component has been subtracted from each point. This results in a vector that is orthogonal to the first principal component, and identifies the direction of the greatest remaining variance in the data. Each subsequent component operates similarly.

Formally, let  $X$  be a  $D \times F$  matrix containing  $D$  data points with  $F$  features each. Assume that  $X$  contains z-scores—that is, columns are zero-mean and unit variance. (If it is not, subtract the sample mean from each row and divide each column by its variance.) PCA identifies an  $F \times F$  matrix  $W$  containing the eigenvectors of the covariance matrix  $X^T X$ . In addition, it identifies a  $D \times F$  diagonal matrix  $\Sigma$  containing the square roots of the eigenvalues of  $X^T X$ .  $W$  and  $\Sigma$  can be arranged so that the eigenvalues appear in descending order, in which case the columns of  $W$  contain eigenvectors in descending order of variance in the data.

This approach was applied to the portion of the SAVE Science dataset relating to a single module. By doing so, initial patterns were uncovered. For example, in a module assessing student understanding of weather patterns, six components were discovered, accounting for 61% of the total variance. Table 4.1 shows the top three components, what it was measuring, and the percent of variance accounted for by that component. The first component encompasses every action a student took while in the weather module. For example, a student could click on several archived newspapers to read headlines and weather reports, ask questions of several different characters, take measurements with a barometer or thermometer, record measurements or other notes to a virtual clipboard, or travel to two other locations via teleportation. In each case, a record would be made in the database of what the student clicked on, said, or recorded, and



where the student traveled. The second component refers to answers a student gave to the end-of-module questions. These questions are varied and include multiple choice, free form, and ranking. All but two of the questions are directly about solving the problem in the environment. The last two are high stakes multiple-choice questions that are assessing the same content as is found in the module. Finally, the third component relates to whether and what data a student chose to graph in the module. For example, typically in the weather module, students graph temperature and barometer readings over time.

**Table 4.1.** Principle component analysis (PCA) results for the weather module (n=146).

Principal Component number	Measured	Total variance accounted for
1	Overall activity	38%
2	Answering questions	8%
3	Graphing activity	5.3%

## Pearson Correlation Coefficients

The next step used in the SAVE Science project was to analyze more closely the trends discovered in the PCA analysis using computed Pearson correlation coefficients and statistical significance tests for a number of pairs of variables. Such tests can indicate whether a pair of variables is statistically non-independent and whether the relationship has enough support in the data that it is highly unlikely for the relationship to appear due to random variations. The analysis cannot determine causality, however.

Statistical significance of the Pearson correlation coefficient is computed using a standard t-test. The SAVE Science study used one-tailed t-tests with an  $\alpha$  level of  $p < 0.05$ . However, for comparisons between a categorical variable (like end-of-module question answers and different teachers) and a numerical variable, unpaired, two-tailed t tests to compare the means of the numerical variables were used. For comparisons between two categorical variables, two-tailed  $\chi^2$  tests were applied. In all cases, the significance level was  $p < 0.05$ .

When applied to the portion of the dataset relating to the same module discussed above, it was found, for example:

- Students who were the most active in the Scientopolis world tended to answer correctly the end-of-module questions that required applying their knowledge of weather to solving the problems in the Scientopolis world.
- However, student activity was negatively correlated with correctly answering questions that do not require knowledge about the world and could be answered with general knowledge. These patterns were backed up by results from the second principal component.
- Correlation analysis also was able to identify significant differences between students taught by different teachers.

If activity in the world is correlated with answering the summary questions at the end as these results suggest, then the researchers' goal of using student actions to predict student understanding of the content is possibly achievable. This goal guided the next data mining steps.

## Cluster Analysis

Cluster analysis, like PCA, has been used extensively over the past several decades. In essence, as its name implies, the point of cluster analysis is to group similar objects together and indicate how closely clusters are related.

For example, the SAVE Science researchers ran the K-means clustering algorithm on the weather dataset. K-means partitions the input data matrix  $X$  into  $k$  partitions  $S = \{S_1, \dots, S_k\}$  in order to minimize the distances between data points within the same partition  $S_i$ . Formally, the algorithm tries to find

$$\arg \min_S \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|$$

where  $\mu_i$  is the mean of data points in partition  $S_i$ . An exact solution to this optimization problem is computationally intractable, so the K-means algorithm uses greedy optimization techniques to find an approximate solution. The results are shown in Table 4.2. This kind of analysis helps visualize the data and identify patterns involving several variables at once. For example, cluster 1 is a cluster of students with a low level of total interactions (“collisions”) in the module compared to other clusters of students, but these students have the highest number of distinct collisions, distinct measurements, and distinct clipboard recordings on average. This might indicate a group of students who are parsimonious in their activity. This can be compared to cluster 2, who have one of the highest levels of interactions in the module but have the lowest number of distinct collision, distinct measurements, and distinct clipboard recordings. It is likely that this group, while active, did not explore enough of the Scientopolis world to make sense of the problem. This is shown by their poor success rate on the end-of-module questions compared to cluster 1. These results are not easily discerned through PCA or correlational analysis.

**Table 4.2.** K-Means cluster analysis of the weather data.

Cluster centroids:							
		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	126	13	14	8	22	27	42
Collisions	186.5159	399.9231	106	286.875	358.8182	145.6296	64.2143
Distinct Collisions	23.4206	24.0769	27.5714	19.25	24.4091	22.1852	22.9048
Measurements	13.2857	13.6923	19.4286	8.875	13.1364	12.7037	12.4048
Distinct Measurements	8.4286	9.6923	11.9286	5.25	7.8636	7.4074	8.4286
Clipboards	7.5556	9	13.0714	4.375	6.5909	7.1111	6.6667
Distinct Clipboards	7.2222	8.9231	12.0714	4.25	6.1364	6.7407	6.5238
Graphs	2.9921	5.7692	8.3571	0.25	2.3636	2	1.8333
Distinct Graphs	1.0476	1.7692	1.9286	0.25	1.0455	1.0741	0.6667
Correct Questions	2.0794	1.1538	3.2143	1.125	1.6364	2.3333	2.2381

## Next Steps

The SAVE Science project is continuing its analysis beyond these steps. The project is currently characterizing students into categories based on some of the above analysis and trying to create models of predictors for success on the end-of-module questions. In this analysis, we are considering whether



predicting for module-based questions differs from using behavior to predict for the high-stakes multiple-choice questions on the same topic.

In the future, it is hoped to explore the impact of time on the analysis. Plans include an interesting visualization: a time-series plot that shows the number of students who still have more interactions to do at each point in time. Do successful students have different patterns of interactions compared to less successful students? Is there a specific point before which behavior has no predictive value for understanding? Using data mining to answer questions such as these will shed light on whether immersive virtual environments can be used to assess aspects of science content and inquiry that are currently not well assessed on typical large-scale assessments.

---

This material is based upon work supported by the National Science Foundation under Grant Nos. DRL 0733286; DGE 0742503; DRL 0814512; DRL 0822308; DRL 1008649; DRL 1157534. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

---

## **IV. SUMMARY**

---



## Summary

This report illustrates the wide range of techniques being applied in projects that seek to create electronic learning environments that are capable of close and constant monitoring of student progress so that the system or the students and teachers can make instructional decisions about what next step will advance the students' learning. What is also evident is that the methods in use for making sense of the rich datasets that are typical of such learning environments come from a range of disciplines. They include methods from statistics (correlations), psychometrics (Item Response Theory), data mining (principal component analysis and cluster analysis), and artificial intelligence (Bayes Nets and Hidden Markov Models). Another pattern from the report is that projects are applying these methods grounded within a framework of knowledge representations based on beliefs about learning in the domains being taught, which brings in approaches from cognitive psychology, education, and cognitive science.

So, what are we to make of this? It points to the need for those interested in measuring student performance in e-learning to be familiar with a range of quantitative techniques from several disciplines. It also indicates that they need an appreciation of cognition and learning to fully understand how the measurement methods can be applied. In short, we need to broaden our expertise beyond the fields within which we were trained so we may familiarize ourselves with related disciplines and the relevant work that informs our understanding of e-learning.

This transformation of the field of educational measurement has already begun as illustrated by the work featured in this report, and it will gather pace as education also transforms to include substantially larger proportions of e-learning as blended models of face-to-face and face-to-technology start to pervade teaching and learning. To ensure that the field of educational measurement is ready to meet this need rather than lagging behind the trend, there are some things that can be done.

### **What can individuals do?**

First, for those of us already working in educational measurement, we can start to educate ourselves and move out of our comfort zones. A good way to start is by attending a conference in an area you are less familiar with. By attending keynotes, plenary sessions, and special interest groups, you can get a sense of what is current in that field, how experts in that domain talk about their subject, and what the opportunities are for future research. Most importantly, try to attend poster sessions, roundtables, and social events like lunches, dinners, or trips during which you can talk with researchers and students from the field. It is through those conversations that deeper understanding and future professional connections can stem over time.

Another way to expand your knowledge is to attend workshops on topics of interest, which are often held at conferences or sometimes independently. A good example is the summer workshops offered at the Pittsburgh Science of Learning Center, in which you can learn about authoring intelligent tutoring systems.

A third approach to broadening your knowledge is to read some journals from the field you want to know more about. This, again, gives you a sense of what work is going on in that area and the research questions that are being asked. If you do such reading before going to a conference in that field, you will have some prior understanding that will help you to identify research you might want to know more about, a particular research group you would like to get to know more about, or an individual whose work you want to make connections to.

The beauty of taking these steps is that you will become an ambassador from your own field who can explain to others what research you are involved in, the courses you teach or take, and how you want to change education for the better. Doing so makes you think critically about your own field and your area of expertise so that you can see how it fits into the larger world of educational research and what similarities, overlaps, and differences there are to other fields.

## What Can Teaching Institutions Do?

Institutions can do a lot to create a new wave of graduates who have the kind of cross-training described here. Some individual students take matters into their own hands and seek out courses in other disciplinary areas at their institutions. At UC Berkeley, for example, a doctoral student from the Quantitative Methods in Education program, who has a major focus in educational measurement, is attending courses on artificial intelligence in the computer science department. At the same time, a graduate student getting a Ph.D. in computer science is enrolled in educational measurement classes in the Quantitative Methods in Education program.

While these enterprising students have found some personal solutions, it is time for institutions to recognize the need and create more opportunities for students wishing to enter the emerging field of measurement in e-learning. A good example of how an institution has done this is the program at Worcester Polytechnic Institute. There students are able to take course sequences that blend the fields of the learning sciences with computer science and EDM. As a first step, institutions could add some core course requirements that ensure that students do some courses in a related discipline. Or they could create joint degree programs in which students do equal parts in different departments.

## What Can Research Funders Do?

The National Science Foundation has already recognized the need for more cross-fertilization among the disciplines to enhance educational research. At a recent NSF- and Organisation of Economic Co-operation and Development-sponsored conference held in Paris, researchers from cognitive science, neuropsychology, computer science, and educational measurement came together to present overviews of the research in their areas. More events like these, for wider audiences, would help to promote such interdisciplinary connections.

## Future Areas of Research

In addition to the directions for future research identified in each of the papers in this report, we discuss here some challenges for the field that we feel could benefit from further research. In the area of knowledge representations, we know that defining domains and representing them is an arduous but necessary task if we are to accurately measure student learning. Future work could investigate ways in which automation of the process of knowledge representation can assist this process, although we recognize that this is not a new idea. Designing systems that would allow content experts to “brain dump” their expertise and then manipulate that into representations that are useful for designing intelligent systems would make the production of learning tools much easier and faster. Such systems would help designers to decide what skills and sub-skills are necessary to measure. Creation of such systems will also require collaboration among cognitive scientists, computer scientists, psychometricians, educators, and content experts.

Another area in which we need to make progress is in creating knowledge representations in more topics and over more domains. Much effort in the field has concentrated on the domains of mathematics and science, but additional research is needed if we are to successfully extend knowledge representation into



other domains (like English literature or history) that don't lend themselves so easily to the kinds of representations we have used in the past.

A further area for growth in the field is methods for measuring in different learning environments, as input devices like cameras (facial and object recognition), graphic pads, e-pens, kinetic movement detectors, natural language processors, haptic devices, etc., widen the types of data we might gather in the learning process. These devices will produce rich streams of information about learners' actions and affect, so future measurement will need to deal with new forms of data and be able to interpret them to make meaningful inferences about the learner. These data sources may also come from less formal educational environments like games, museum exhibits, and blended learning environments that mix hands-on and computer-based activities in new forms of augmented reality.

The challenge for the field of educational measurement will be ensuring that in whatever direction learning systems move, that the measurement systems produce outputs that are relevant to the needs of learners and teachers throughout the learning transactions during and at the conclusion of the learning sequence. This means that we will focus on determining the optimal grain sizes to collect and aggregate data, how best to provide feedback to learners, and what kinds of information needs to be accessible to teachers to allow them to support learners. Systems will not be used in the classroom if they do not integrate with teachers' ways of teaching.

An emerging area for educational measurement is *educational analytics*, in which the large datasets that are produced from e-learning are being data mined to reveal patterns in learning behaviors. The focus in educational analytics is at a larger grain size that might span a whole course or even several courses, rather than the individual steps in learning that are the focus of the work described in this report. These kinds of analyses are often focused on larger-grain-size questions like, How can we predict which students are at risk of dropping out or are in need of additional support to increase their success, and confidence, in the learning process? This is of particular interest in higher education where more and more learning is through electronic media, thereby putting some distance between a learner and a teacher. Educational analytics might monitor such data as students' participation in an online forum, their level of interaction with other students or the teacher, and their use of online resources for the class.

In summary, as education inevitably incorporates more and more technology into the learning enterprise, the field of educational measurement needs to play its part in making sure that the knowledge representations, the data collected, the analyses performed, and the feedback delivered to learners and educators all coordinate in intelligent learning systems that personalize and optimize learning for all.



---

## V. REFERENCES

---



## References

- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R., Reder, L. M., & Simon, H. A. (1998). Radical constructivism and cognitive psychology. In D. Ravish (Ed.), *Brookings papers on educational policy: 1998* (pp. 384). Washington, DC: The Brookings Institution Press.
- Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5). Mahwah, NJ: Lawrence Erlbaum Associates.
- Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? . In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143–166). Maple Grove, MN: JAM Press.
- Baker, S.J.D.R., & Yacef, Y. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17.
- Bangalore, S., Di Fabbriozio, G., & Stent, A. (2008). Learning the structure of task-driven human-human dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7), 1249–1259.
- Barnes, T., & Stamper, J. (2010). Automatic hint generation for logic proof tutoring using historical data. *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems*, Pittsburgh, PA. 3–14.
- Beal, C., Mitra, S., and Cohen, P.R. (2007). Modeling learning patterns of students with a tutoring system using Hidden Markov Models. *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 238–245.
- Behrens, J. T., Frezzo, D., Mislevy, R., Kroopnick, M., & Wise, D. (2008). Structural, Functional, and Semiotic Symmetries in Simulation-Based Games and Assessments. In E. Baker, J. Dickieson, W. Wufleck & H. F. O’Neil (Eds.), *Assessment of Problem Solving Using Simulations* (pp. 59–80). New York: Lawrence Erlbaum Associates.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Borgman, C. L., Abelson, H., Johnson, R., Koedinger, K. R., Linn, M. C., Lynch, C. A., . . . Szalay, A. (2008). *Fostering learning in the networked world: The cyberlearning opportunity and challenge: A 21st century agenda for the National Science Foundation*. Arlington, VA: National Science Foundation. Retrieved from [http://www.nsf.gov/pubs/2008/nsf08204/nsf08204.pdf?govDel=USNSF\\_124](http://www.nsf.gov/pubs/2008/nsf08204/nsf08204.pdf?govDel=USNSF_124)
- Boyer, K. E., Ha, E. Y., Wallis, M. D., Phillips, R., Vouk, M. A., & Lester, J. C. (2009). Discovering tutorial dialogue strategies with hidden Markov models. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, Brighton, U. K. 141–148.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Buckley, B.C., Gobert, J., Horwitz, P., & O’Dwyer, L. (2010). Looking inside the black box: Assessments and decision-making in BioLogica. *International Journal of Learning Technology*, 5(2), 166-190.
- Cade, W., Copeland, J., Person, N., & D’Mello, S. (2008). Dialog modes in expert tutoring. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, Montreal, Canada. 470–479.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control variables strategy. *Child Development*, 70(5), 1098-1120.
- Chi, M., Jordan, P., VanLehn, K., & Litman, D. (2009). To elicit or to tell: Does it matter? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 197–204.

- Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25(4), 471–533.
- Chipman, S. F., Schraagen, J. M., & Shalin, V. L. (2000). Introduction and history. In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.), *Cognitive task analysis* (pp. 3–23). Mahwah, NJ: Lawrence Erlbaum Associates.
- Clark, R. E., Feldon, D. F., Merriënboer, J. J. g., Yates, K. A., & Early, S. (2008). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. v. Merriënboer, & M. R. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577-593). New York: Lawrence Erlbaum Associates.
- Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach*. New York: Routledge.
- Clements, D. H., Sarama, J., & Liu, X. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-based Early Maths Assessment. *Educational Psychology*, 28(4), 457-482.
- Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *TEAM—Tools for early assessment in mathematics*. Columbus, OH: McGraw-Hill Education.
- Committee on Conceptual Framework for the New K–12 Science Education Standards, & National Research Council. (2011). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas* (pp. 320). Washington, DC: National Academies Press.
- Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12, 371-417.
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4, 253–278.
- Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1993). The predictive validity of student modeling in the ACT programming tutor. In P. Brna, S. Ohlsson, & H. Pain (Eds.), *The Proceedings of AI ED 93* (pp. 457–464). Charlottesville, VA: Association for the Advancement of Computing in Education.
- Corbin, J., & Strauss, A. L. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1), 3–21.
- de Lemos, M., & Doig, B. (1999). *Who am I? Developmental assessment manual*. Camberwell, Victoria, Australia: Australian Council for Educational Research.
- Doig, B., & de Lemos, M. (2000). *I can do maths*. Camberwell, Victoria, Australia: Australian Council for Educational Research.
- Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17, 397-434.
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fadel, C., Honey, M., & Pasnick, S. (2007). Assessment in the age of innovation. *Education Week*, 26(3), 4-40.
- Feigenbaum, E.A., & McCorduck, P. (1983). *The fifth generation: Artificial intelligence and Japan's computer challenge to the world*. Reading, MA: Addison-Wesley.
- Forbes-Riley, K., & Litman, D. (2009). Adapting to student uncertainty improves tutoring dialogues. *Proceedings of the 14<sup>th</sup> International Conference on Artificial Intelligence in Education*.
- Fossati, D., Di Eugenio, B., Brown, C., Ohlsson, S., Cosejo, D., & Chen, L. (2009). Supporting computer science curriculum: Exploring and learning linked lists with iList. *IEEE Transactions on Learning Technologies*, 2(2), 107–120.
- Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., & Chen, L. (2010). Generating proactive feedback to help students stay on track. *Proceedings of the 10<sup>th</sup> International Conference on Intelligent Tutoring Systems*, 315-317.
- Fox, B. A. (1993). *The human tutorial dialogue project*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Gentner, D. (1980). *The structure of analogical models in science (report No. 4451, NTIS No. AD-A087-625)*. Springfield, VA: National Technical Information Service, U.S. Department of Commerce.



- Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gobert, J. (2005). The effects of different learning tasks on conceptual understanding in science: teasing out representational modality of diagramming versus explaining. *Journal of Geoscience Education*, 53(4), 444–455.
- Gobert, J.D., Pallant, A.R., & Daniels, J.T.M. (2010). Unpacking inquiry skills from content knowledge in geoscience: a research and development study with implications for assessment design. *International Journal of Learning Technology*, 5(3), 310-334.
- Gobert, J., Sao Pedro, M., Baker, R., Toto, E., & Montalvo, O. (accepted). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. To appear in the *Journal of Educational Data Mining*.
- Graesser, A.C., Person, N.K., & Magliano, J.P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9, 495-522.
- Hand, D.J., Blunt, G., Kelly, M. G. & Adams, N. M. (2000). Data mining for fun and profit, with discussion. *Statistical Science* 15, 111–131.
- Hitt, B. (2012). *Knowledge discovery from Selene data*. Paper presented at the American Educational Research Association Conference, Vancouver.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping with constraint satisfaction. *Cognitive Science*, 13, 295–355.
- Huizinga, J. (1950). *Homo ludens: A study of play-element in culture* (Paperback ed.). Boston: Beacon Press.
- Jeong, H., Gupta, A., Roscoe, R., Wagster, J. Biswas, G., & Schwartz, D. (2008). Using hidden markov models to characterize student behaviors in learning-by-teaching environments. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 614–625.
- Kersey, C., Di Eugenio, B., Jordan, P., & Katz, S. (2009). KSC-PaL: A peer learning agent that encourages students to take the initiative. *Proceedings of the NCAAL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, Boulder, CO. 55-63.
- Ketelhut, D. J., Nelson, B., Schifter, C., & Kim, Y. (2010). Using immersive virtual environments to assess science content understanding: the impact of context. In Kinshuk, D. G. Sampson, J. M. Spector, P. Isaías, D. Ifenthaler and R. Vasiu (Eds.), *Proceedings of the Iadis International Conference on Cognition and Exploratory Learning in the Digital Age, Celda2010* (pp. 227–230). Timisoara, Romania.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1-48.
- Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211-228.
- 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., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). New York, NY: Cambridge University Press.
- Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*.
- Koedinger, K., & Nathan, M. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences*, 13(2), 129-164.
- Koedinger, K., Suthers, D., & Forbus, K. (1999). Component-based construction of a science learning space. *International Journal of Artificial Intelligence in Education*, 10, 292-313.
- Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

- Langhoff, S., Cowan-Sharp, J., Dodson, E., Damer, B., Ketner, B., & Reese, D. D. (2009). *Workshop report: Virtual worlds and immersive environments*. (NASA/CP-2009-214598). Moffett Field, CA: NASA Ames Research Center. Retrieved from [http://event.arc.nasa.gov/main/home/reports/CP-2009-214598\\_Virt\\_Worlds.pdf](http://event.arc.nasa.gov/main/home/reports/CP-2009-214598_Virt_Worlds.pdf)
- Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. L. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie, & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 75–105). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Linacre, J. M. (2005). *A user's guide to Winsteps/Ministep Rasch-model computer program*. Chicago IL: Winsteps.com.
- Martin, J., & VanLehn, K. (1995). Student assessment using Bayesian nets. *International Journal of Human-Computer Studies*, 42, 575-591.
- Merrill, M. D. (2002). First principles of instruction. *Educational Technology, Research, & Development*, 50(3), 43–59.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to Evidence-Centered Design*. Retrieved from <http://www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf>
- Mislevy, R.J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2002). *Design Patterns for Assessing Science Inquiry*. Unpublished Manuscript, Washington, DC.
- Mislevy, R., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., & Haertel, G. (2003). *Design patterns for assessing science inquiry*. Menlo Park, CA: SRI International.
- Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in cognitive diagnosis. *User Modeling and User-Adopted Interaction*, 5, 253-282.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.
- National Research Council. (1996). *National Science Education Standards*. Washington, DC: The National Academies Press.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council. (2011). *Learning science through computer games and simulations*. Washington, DC: National Academies Press.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X., & Kershaw, T. C. (2007). Beyond the code-and-count analysis of tutoring dialogues. *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 349–356.
- Pardos, Z. A., & Heffernan, N. T. (2010). *Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm*. Paper presented at the 3rd International Conference on Educational Data Mining, Pittsburgh, PA.
- Pardos, Z. A., Heffernan, N. T., Anderson, B. S., & Heffernan, C. L. (2010). Using fine-grained skill models to fit student performance with Bayesian Networks. In C. Romero, S. Ventura, M. Pechenizkiy, & R.S.J.D. Baker (Eds.), *Handbook of Educational Data Mining* (pp. 417-426). Boca Raton, FL: CRC Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Quellmalz, E. S., Silberglitt, M. D., & Timms, M. J. (2011). *How can simulations be components of balanced state science assessment systems?* San Francisco: WestEd.



- Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2012). 21st century dynamic assessment. In M. Mayrath., J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.
- Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (2009). *Assessment of student learning in science simulations and games*. Paper commissioned for the National Research Council Workshop on gaming and simulations, October 6–7. Washington, DC. Retrieved from [http://www7.nationalacademies.org/bose/Schneider\\_Gaming\\_CommissionedPaper.pdf](http://www7.nationalacademies.org/bose/Schneider_Gaming_CommissionedPaper.pdf)
- Quellmalz, E., Timms, M., Silberglitt, M., & Buckley, B. (2012) Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research on Science Teaching*, 49(3), 363-393.
- Quinn, J., & Alessi, S. (1994). The effects of simulation complexity and hypothesis-generation strategy on learning. *Journal of Research on Computing in Education*, 27(1), 75-91.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
- Reese, D. D. (2009). Structure mapping theory as a formalism for instructional game design and assessment. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *New frontiers in analogy research: Proceedings of the 2nd International Conference on Analogy (Analogy '09)* (pp. 394–403). Sofia, Bulgaria: New Bulgarian University Press.
- Reese, D. D. (2010). Introducing flowometer: A CyGaMEs assessment suite tool. In R. V. Eck (Ed.), *Gaming & cognition: Theories and perspectives from the learning sciences* (pp. 227–254). Hershey, PA: IGI Global.
- Reese, D. D., Seward, R. J., Tabachnick, B. G., Hitt, B., Harrison, A., & McFarland, L. (in press). Timed Report measures learning: Game-based embedded assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York: Springer.
- Reese, D. D., & Tabachnick, B. G. (2010). The moment of learning: Quantitative analysis of exemplar gameplay supports CyGaMEs approach to embedded assessment. In J. Earle (Ed.), *Building a knowledge base to inform educational practice in STEM: Examples from the REESE portfolio*. Symposium conducted at the Society for Research on Educational Effectiveness 2010 Annual Research Conference, Washington, DC. Structured abstract retrieved from <http://www.sree.org/conferences/2010/program/abstracts/191.pdf>
- Ritter, S., Anderson, J. R., Cytrynowicz, M., & Medvedeva, O. (1998). Authoring content in the PAT Algebra Tutor. *Journal of Interactive Media in Education*, 98(9).
- Romero, C., & Ventura, S. (2007). Educational data mining: a survey from 1995-2005. *Expert System with Applications*, 33(1), 135-146.
- Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. D. (Eds.).(2011). *Handbook of educational data mining*. Boca Raton: CRC Press.
- Rowe, J. P., & Lester, J. C. (2010). Modeling user knowledge with dynamic bayesian networks in interactive narrative environments. *Proceedings of the 6<sup>th</sup> AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: theory, methods and applications*. New York: The Guilford Press.
- Russell, S., & Norvig, P. (2010). *Artificial Intelligence; A modern approach* (Third ed.). Upper Saddle River, New Jersey: Prentice Hall.
- Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (in press). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict the transfer of inquiry skill. *User Modeling and User-Adapted Interaction*.
- Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York: Routledge.

- Shavelson, R., Wiley, E. W., & Ruiz-Primo, M. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, 36(1), 61–71.
- Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73-92). Maple Grove, MN: JAM Press.
- Soller, A. & Stevens, R. (2007). Applications of stochastic analyses for collaborative learning and cognitive assessment. *Advances in Latent Variable Mixture Models*, in G. R. Hancock and K. M. Samuelsen, Eds. Information Age Publishing, 217–253.
- Stamper, J. C., & Koedinger, K. R. (2011). Human-machine student model discovery and improvement using DataShop. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in education: 15th International Conference, AIED 2011* (pp. 353–360). Auckland, New Zealand.
- Strauss, A. L., & Corbin, J. (1991). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park: Sage Publications.
- Strauss, A. L., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Thousand Oaks: Sage Publications.
- Studer, R., Benjamins, V.R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data and Knowledge Engineering*, 25(1-2), 161-197.
- Tatsuoka, C. (1996). *Sequential classification on partially ordered sets*. Doctoral dissertation, Cornell University.
- Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Applied Statistics*, 51, 337–350.
- Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society, Series B*, 143-157.
- Tatsuoka, K. K. (2009). *Cognitive assessment : An introduction to the rule space method*. New York, NY: Routledge.
- Tetreault, J. R., & Litman, D. J. (2008). A reinforcement learning approach to evaluating state representations in spoken dialogue systems. *Speech Communication*, 50(8–9), 683–696.
- Thomson, S., Rowe, K., Underwood, C., & Peck, R. (2005). *Numeracy in the early years: Project Good Start*. Camberwell, Victoria, Australia: Australian Council for Educational Research.
- Timms, M. (2007). Using Item Response Theory (IRT) in an Intelligent Tutoring System. Proceedings of the Artificial Intelligence in Education 2007 Conference, Marina Del Ray, CA. IOS Press, Washington DC. *Frontiers in Artificial Intelligence and Applications*. Vol. 158 (213–221).
- U.S. Department of Education Office of Educational Technology. (2010). *Transforming American education: Learning powered by technology: National education technology plan 2010*. Washington, DC.
- Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning*, 9, 83–130.
- What Works Clearinghouse. (2009). *WWC Intervention report: Cognitive Tutor<sup>®</sup> Algebra I*. Washington, DC: Institute of Education Sciences. Retrieved from [http://ies.ed.gov/ncee/wwc/pdf/intervention\\_reports/wwc\\_cogtutor\\_072809.pdf](http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_cogtutor_072809.pdf)
- What Works Clearinghouse. (2010). *WWC Intervention report: Carnegie learning curricula and Cognitive Tutor<sup>®</sup> software*. Washington, DC: Institute of Education Sciences. Retrieved from [http://ies.ed.gov/ncee/wwc/pdf/intervention\\_reports/wwc\\_cogtutor\\_083110.pdf](http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_cogtutor_083110.pdf)
- Williamson, D., Mislevy, R., & Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716–730.



- Woolf, B. P. (2009). *Building Intelligent Interactive Tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, MA: Morgan Kaufmann.
- Wright, W. (2003). *Dynamics for designers*. Keynote session presented at the 2003 Game Developers Conference, San Jose, CA. [http://www.gamasutra.com/features/gdcarchive/2003/Wright\\_Will.ppt](http://www.gamasutra.com/features/gdcarchive/2003/Wright_Will.ppt)
- Wright, W. (2006a). Dream machines. *Wired*, 14(4). Retrieved from <http://www.wired.com/wired/archive/14.04/wright.html>
- Wright, W. (2006b). *What's next in game design*. Keynote Session presented at the 2006 Game Developer's Conference, Game Developer's Conference TV, San Jose, CA. [http://www.gamasutra.com/features/20060324/sanchez\\_01.shtml](http://www.gamasutra.com/features/20060324/sanchez_01.shtml)
- Wright, W. (2009). *The gaming paradigm*. Keynote session presented at the 5th annual Innovations in e-Learning Symposium, George Mason University. [http://www.google.com/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=3&ved=0CDYQtwIwAg&url=http%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DxrxBzn6M5OA&ei=YaoUT8TEEO\\_q0QGptYmZAw&usg=AFQjCNHTHZrk1rXPndOsBvStgp0aCbgMiQ](http://www.google.com/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=3&ved=0CDYQtwIwAg&url=http%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DxrxBzn6M5OA&ei=YaoUT8TEEO_q0QGptYmZAw&usg=AFQjCNHTHZrk1rXPndOsBvStgp0aCbgMiQ)
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.