Leveraging Educational Data Mining for Real-time Performance Assessment of Scientific Inquiry Skills within Microworlds

JANICE D. GOBERT, MICHAEL A. SAO PEDRO, RYAN S.J.D. BAKER, ERMAL TOTO, and ORLANDO MONTALVO
Department of Social Sciences and Policy Studies
Learning Sciences and Technologies Program
Worcester Polytechnic Institute

_____

We present *Science Assistments*, an interactive environment, which assesses students' inquiry skills as they engage in inquiry using science microworlds. We frame our variables, tasks, assessments, and methods of analyzing data in terms of *evidence-centered design*. Specifically, we focus on the *student model*, the *task model*, and the *evidence model* in the *conceptual assessment framework*. In order to support both assessment and the provision of scaffolding, the environment makes inferences about student inquiry skills using models developed through a combination of text replay tagging [cf. Sao Pedro et al. 2011], a method for rapid manual coding of student log files, and *educational data mining*. Models were developed for multiple inquiry skills, with particular focus on detecting if students are testing their articulated hypotheses, and if they are designing controlled experiments. Student-level cross-validation was applied to validate that this approach can automatically and accurately identify these inquiry skills for new students. The resulting detectors also can be applied at run-time to drive scaffolding intervention.

Keywords: Performance assessment of inquiry skills, educational data mining, machine learning, text replay tagging

_____

Author's addresses: Department of Social Sciences and Policy Studies, Learning Sciences and Technologies Program, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, 01609, {jgobert, mikesp, rsbaker, toto, amontalvo}@wpi.edu

1. INTRODUCTION

It is well-acknowledged that in order to be competitive in 21st century workplaces, students need to understand science deeply, and possess well-honed learning strategies that will allow them to apply their science knowledge in flexible ways. For example, the national frameworks for science emphasize inquiry skills [NRC 2011], also referred to as 'scientific practices' by the *National Assessment of Educational Progress* [Champagne et al. 2004], claiming that these skills are critical to science reform efforts. As such, efforts have been made to incorporate inquiry within the classroom and to find ways to address the assessment needs of today's schools.

In regards to assessment, there are two commonly used approaches for doing so, *short answer tests* of specific inquiry skills [cf. Alonzo and Aschbacher 2004] and *hands-on performance assessments* [cf. Gotwals and Songer 2006]. Short answer tests can be incorporated into large-scale standardized assessments, but it is unclear whether these properly identify inquiry skills [Black 1999; Pellegrino et al. 2001], or whether they are well aligned to current national frameworks [Quellmalz et al. 2007].

For example, Scalise et al. [2009] studied 77 simulation products for middle and high school, and in more than half of these projects, conventional, paper-and-pencil pre- versus post-test comparisons were used as assessments. This type of assessment has relatively limited possibilities for measuring the complex science knowledge and skills that their instruction was designed to target [Quellmalz et al. 2009]. Hands on performance assessments are more authentic because they require specific skills to solve real problems [Baxter and Shavelson 1994; Ruiz-Primo and Shavelson 1996]. However, these are seldom used in schools, largely due to the difficulty of reliable administration and the resulting high cost.

Another critical issue is that since inquiry skills are developed in rich scientific contexts, it is crucial that they also be assessed within the contexts in which they are learned and embedded [Mislevy et al. 2003]. Lastly, in order to yield reliable assessments, a great deal of data is required; this is yet another problem that has hampered the development of inquiry assessment measures, methods, and procedures [Shavelson et al. 1999]. Thus, despite the acknowledged need for science inquiry skills, typical classrooms often focus on rote learning of vocabulary, facts, and formulas, due to the difficulty of implementing inquiry [de Jong et al. 2005], and

difficulty assessing science process skills [Fadel et al. 2007; Mislevy et al. 2003; Shavelson et al. 1999].

In response to these implementation and assessment difficulties, we are developing a web-based learning environment, *Science Assistments* [Gobert, Heffernan, Ruiz & Ryung, 2007; Gobert, Heffernan, Koedinger & Beck, 2009] to provide scalable, reliable performance-based assessment of authentic inquiry skills. This interactive environment assesses inquiry skills as students experiment within science microworlds [Papert 1980], computerized models of real-world phenomena whose properties can be inspected and changed [Pea and Kurland 1984; Resnick 1997].

As part of this approach, we utilized the *conceptual assessment framework* (CAF) of the *evidence-centered design* (ECD) framework [e.g., Mislevy et al. 2003; Mislevy et al., this issue], because we perceived it to be most important to our work on inquiry assessment. We describe here our *student model* (the inquiry skills we aim to measure), *task model* (the activities designed to elicit demonstration of skill), and *evidence model* (our approach to assessing and tracking student knowledge). We also focus on how we leveraged techniques from the *educational data mining* (EDM) literature [cf. Baker and Yacef 2009; Romero and Ventura 2010] to develop our *evidence model*s, and reliably assess inquiry process skills associated with data collection [Sao Pedro et al., 2011].

The remainder of this paper is organized as follows. First, we provide background literature on ECD and prior work on assessing inquiry skills within computerized environments. Then, we present an overview of *Science Assistments*. Next, we present our realization of the CAF emphasizing our EDM-based approach to developing *evidence model*s of two process skills associated with data collection. Finally, we present our conclusions and future work.

2. BACKGROUND

2.1 Evidence-Centered Design

ECD, a cognitively principled framework for designing assessments, suggests that one should relate the learning to be assessed, as specified in a *student model*, to a *task model* that specifies features of tasks and questions that would elicit the evidence of learning, then to an *evidence model* that specifies the quality of student responses that would indicate levels of proficiency

[Messick 1994; Mislevy et al. 2003; Mislevy et al., this issue; Pellegrino et al. 2001]. In more recent documents on ECD, its authors have added the presentation model and the assembly model as components to their assessment framework [Mislevy et al. 2006] as well as descriptions of the computational processes that tie them together in practice, which are known as the *four-process model* [Mislevy et al., this issue]; however, we did not utilize these components in our work.

For our *student model*, we defined a set of inquiry skills and sub-skills that have been identified as being pedagogically-relevant to the assessment of inquiry [NRC 1996, 2011]. The *task model* refers to the specification of the microworlds and activities conducted in the microworld that can reveal students' proficiencies for each skill of interest. Our *evidence model* is comprised of two sources of student data. The first is the work products students generate while conducting inquiry and the second are the work processes, namely the types of actions / behaviors they engage in while experimenting during inquiry. These are then aggregated and analyzed to yield performances that are used to identify evidence of students' proficiencies for each skill of interest [NRC 1996, 2011]. Further, the work products and processes of inquiry that result are interpreted as evidence of proficiency on the skills of interest.

2.2 Inquiry Assessment in the Context of Science Simulations

Previously in the paper, we noted problems with traditional assessments of inquiry skills. Rich learning environments, which generate large streams of log data, are potentially of great use for performance assessment; however, these learning environments often are not used for performance assessment to the extent they could [Scalise et al. 2009] in order to measure the complex science knowledge and skills that their instruction was designed to target [Quellmalz et al. 2009].

There are several reasons why this is the case. First, the development of assessment methods for simulation-based learning is lagging behind in terms of theoretical grounding in the learning and assessment literature [Quellmalz et al. 2009]. Second is the lack of theoretically motivated guiding principles upon which to parse, aggregate, and analyze the huge streams of log data that are generated as students learn with simulations. Third, specifically, the skills being assessed are psychologically multidimensional and as such, traditional psychometric methods such as

classical test theory (CTT) [e.g., Crocker and Algina 2006] and *unidimensional item response theory* (UIRT) [de Ayala 2009; Hambleton and Jones 1993] cannot appropriately model their complexity. This, according to Quellmalz et al. [2009], is principally due to the complexity of the simulations, and is summarized by their following four main characteristics [Williamson et al. 2006; see also Rupp et al. 2010]:

1) the task requires the learner to complete many, non-trivial, domain-relevant steps/processes,

2) multiple elements or features for each task are captured and considered in the determination of skills for assessment purposes or diagnostic feedback,

3) there is potential for wide variability in the data for each task, and

4) the task as a whole has many components that are not independent from each other.

Though these complexities exist, several researchers are attempting to overcome these difficulties by employing a variety of tactics. We discuss some these in the next section.

2.3 Prior Approaches to Assessment for Simulation-Based Learning

Noting the limitations of CTT and UIRT as methods for performance-based assessments, other methods have been applied to data in order to model students' knowledge and skills learned in simulations and similarly complex learning environments. These methods fit largely into two categories, (1) models developed through knowledge engineering and/or cognitive task analysis, and (2) models developed through data mining and/or machine learning methods. We describe these in more detail below.

2.3.1 Knowledge Engineering / Cognitive Task Analysis Approaches. In *knowledge engineering/ cognitive task analysis* approaches, rules are defined a priori that encapsulate specific behaviors [Koedinger et al. 1998; McElhaney and Linn 2008, 2010] or differing levels of systematic experimentation skill [Buckley et al. 2006; Buckley et al. 2010]. For example, Schunn and Anderson [1998] engineered a rule-based *Adaptive Control of Thought—Rational* (ACT-R) model of scientific inquiry based on an assessment of skill differences between experts and novices on formulating hypotheses, exploring, analyzing data, and generating conclusions

[e.g., Anderson and Lebiere 1998]. Briefly, ACT-R evolved from earlier theories of human cognition and models of cognitive architecture [Newell 1990; Newell and Simon 1972] and is used to model "overt, observable human behavior(s)" [Anderson and Lebiere 1998, p. 10]. ACT-R describes cognition as involving declarative knowledge (i.e., knowledge *about* things), and procedural knowledge (i.e., *skills that act on knowledge*); procedural knowledge is implemented in ACT-R models as production rules.

With ACT-R in mind, *knowledge-engineering models* can be leveraged using a method called *model-tracing*, where student responses are matched to a knowledge-engineered cognitive model of expert/correct behavior that includes declarative knowledge and production rules and, in some cases, specific misconceptions termed "bugs" [Koedinger et al. 1997]. Model-tracing has been applied in the domain of scientific inquiry in a separate study by Gobert and Koedinger [2011], who used this approach with production rules to auto-score students' inquiry on the use of the *control-of-variables* (CVS) strategy [cf. Chen and Klahr 1999], where all but the target variable is changed across trials within a science microworld.

Model-tracing assessments and other approaches can be, in turn, utilized within *knowledge-tracing* [Corbett and Anderson 1995], a method for assessing latent knowledge from correct and incorrect performance. Reye [2004] has shown that knowledge-tracing models are a simple form of *Bayesian Networks / Bayes nets* (BNs) [e.g., Almond et al. in press]. More complex BNs have also been used in assessing student knowledge in science. For example, Martin and Van Lehn [1995] used BNs to assess procedural knowledge for physics within the *Andes* learning environment. Rowe and Lester [2010] developed dynamic BN models of middle school students' narrative, strategic, and curricular knowledge as students they explored within a 3D immersive environment on microbiology, Crystal Island. Rupp et al. (this issue) use Bayes nets and related diagnostic measurement tools to model multivariate skill profiles for network engineering based on performance in an interactive digital learning environment.

2.3.2 Educational Data Mining / Machine Learning Approaches. In *EDM / machine learning approaches* [cf. Baker and Yacef 2009; Romero and Ventura 2010], student inquiry behaviors are discovered from data. For example, Stevens et al. [2004] used a self-organizing *artificial neural network* [Bryson and Ho 1969; cf. Russell and Norvig 2009] to build models of novice

and expert performance using transition logs within the HAZMAT high school chemistry learning environment. They then leveraged those models to construct a *hidden Markov model* [Baum and Petrie 1966; cf. Russell and Norvig 2009] for identifying learner trajectories through a series of activities.

Similar approaches have also been used to distinguish students' problem solving strategies within exploratory learning environments. For example, Bernardini and Conati [2010] used *clustering techniques* [cf. Witten and Frank 2005] and *class association rules* [Liu et al. 1998; Agrawal and Srikant 1994] to capture learner models of effective and ineffective learning strategies within an exploratory learning environment for learning about a constraint satisfaction algorithm. Ghazarian and Noorhosseini [2010] constructed task-dependent and task-independent machine-learned models to predict skill proficiency in computer desktop applications.

Our work on assessing student inquiry skills in *Science Assistments* (www.inq-its.org) uses techniques from EDM but builds off of a different history of EDM methods, namely the work to develop automated detectors of student behaviors, specifically disengagement. In this line of research, Walonoski and Heffernan [2006] as well as Baker et al. [2008b] successfully built and validated detectors of "gaming the system" (whereby students simply try to get as many hints as possible rather than learn the material) by triangulating qualitative field observations with features gleaned from log files. Detectors of gaming the system have also been developed using a training set labeled using text replays, a method for quickly annotating log files by hand [Baker and de Carvalho 2008; Baker et al. 2010].

As was done in this work, we utilize text replays to annotate log data, and then extrapolate models from the data that can label sequences of student behavior in terms of inquiry performance. Lastly, we use knowledge-engineered models to assess students' work products. Later in this paper, we provide more details on our methods and give evidence on their reliability and validity. In the next section, we provide a description of our learning environment and its components.

3. OVERVIEW OF SCIENCE ASSISTMENTS

Leveraging infrastructure development from other funded projects [Gobert et al. 2007; Gobert et al. 2009; Gobert and Baker 2010] and other prior work by the *Science Assistments* project team as well as current research findings and recent calls for performance assessment [Fadel et al. 2007; Quellmalz et al. 2007], we have developed an approach to science inquiry skills assessment that affords the rigor and validity of performance tests and the simplicity of large-scale assessments.

The *Science Assistments* system, significantly adapted from the original *ASSISTments* system [Razzaq et al. 2005; Heffernan et al. 2006] for mathematics, aims to enable automatic assessment of inquiry skills and provide real-time support for students as they engage in inquiry using interactive microworlds across several topics within physical, life, and earth science [Gobert et al. 2007; Gobert et al. 2009].

Science Assistments enables a moderate degree of student control, less than in purely exploratory learning environments [Amershi and Conati 2009], but more than in classic model-tracing tutors [Koedinger and Corbett 2006] or constraint-based tutors [Mitrovic et al. 2001]. *Science Assistments* also differs from other microworld-based discovery environments [e.g., White and Frederiksen 1998; van Joolingen and de Jong 2003; Buckley, Gobert, & Horwitz 2006; McElhaney and Linn 2008, 2010; Buckley et al. 2010; de Jong et al. 2010] in several ways.

First, it was designed to prioritize the assessment of inquiry skills, rather than the learning of science, which many other inquiry systems emphasize (although content can be learned in our system as one engages in inquiry, in fact, this is considered optimal for science learning). Second, it aims to track acquisition of inquiry skills across domains. Finally, our system aims to scaffold students' inquiry skills in real-time as students conduct investigations within microworlds.

This scaffolding design lead to a guided system; that is, we aimed to strike a balance between *open-ended* and *guided inquiry* [cf., Kirschner et al. 2006; Hmelo-Silver et al. 2007] such that students can be assessed on inquiry skills and so that they are provided with support so they do not flounder or engage in haphazard inquiry [Buckley et al. 2006]. For example, a student might

begin by exploring a microworld, go on to make a hypothesis, return to exploring some more, then make another hypothesis, and test it.

When asked to warrant his/her claim, the student might realize that they did not collect enough data, and then return to the experiment phase again. Thus, students may take various valid paths to finish an inquiry task, so our system needed to be flexible enough to track what each student is doing during each phase in order to provide assessment data to the teacher, as well as real time tutoring to the student.

In terms of the context in which our data are collected, it is important to note that students come to our inquiry environment having *first* studied a science topic in their regular science class; thus, we decided to use our environment for assessment of these previously learned inquiry skills. If we were to design an inquiry environment expressly for learning, we likely would design a system with substantially more open-ended exploration, and it is not difficult to imagine that such an environment would not suffice as an assessment environment.

Thus, because our environment is considered first an assessment environment, rather than a pure learning environment, we have chosen to operationalize inquiry skills so as to support their auto-assessment. As such, we focus on defining and, in turn, assessing skills of inquiry, namely, hypothesizing, conducting experiments, analyzing data, and warranting claims, and their respective sub-skills; a similar approach was taken in the original development of intelligent tutoring systems for math [Koedinger and Corbett 2006].

Rather than including larger conceptual knowledge that pertains to inquiry such as how theory and disciplinary knowledge enter into the inquiry process, we have chosen to focus on the assessment of the more well-defined inquiry skills, as described above. These larger conceptual issues, we feel, are beyond the scope of an inquiry environment and possibly beyond the scope of middle school science instruction as well.

Finally, given the plethora of studies that have evidenced students' difficulties in executing inquiry, including difficulties with *hypothesis-formation* [Chinn and Brewer 1993; Klahr and Dunbar 1988; Kuhn et al. 1995; Njoo and de Jong 1993; van Joolingen and de Jong 1997; Glaser et al. 1992], difficulties with *conducting experiments* [Glaser et al. 1992; Reimann 1991; Tsirgi 1980; Shute and Glaser 1990; Kuhn 2005; Schunn and Anderson 1998, 1999; Harrison and

Schunn 2004; McElhaney and Linn 2008, 2010); difficulties with *interpreting data*, and *linking hypotheses and data* [Chinn and Brewer 1993; Klahr and Dunbar 1988; Kuhn et al. 1995], and difficulties with *communicating findings* [Krajcik et al. 1998; McNeill and Krajcik 2007], it seems reasonable to us that an environment like ours, which as designed to assess and assist with both teachers and students as integral players, is critically what is needed at this time.

There are many advantages of our system for educational research, instruction, assessment, and adaptive scaffolding, and scalability as follows. In terms of research, we can effortlessly and accurately monitor and record every student action, affording thorough analysis of student learning and performance. And because our assessment and instruction are seamlessly integrated, skills are developed and assessed in rich contexts in which they are developing [Mislevy et al. 2003], additional class time is not needed for assessment, and formative assessment data is provided to teachers about his/her students on their science process skills and sub-skills.

Regarding adaptive instruction, we can diagnose at risk students early and monitor these students using our assessment data to drive adaptive scaffolding via our pedagogical agent, Rex, a cartoon dinosaur who provides individualized feedback [Gobert and Baker 2012]. With regard to scalability, because these materials are web-based, virtually any teacher with access to the web can use our materials.

3.1 Components of Science Assistments: Microworlds & Inquiry Support Widgets

Important to inquiry are our microworlds of scientific phenomena for *Physical, Life, and Earth Science*, which are implemented using the *Open Laszlo* platform (www.openlaszlo.org) and accompanying *lzx* language. As previously stated, a microworld [Papert 1980] is a runnable, computerized model of real-world phenomena whose properties can be inspected and changed [Pea and Kurland 1984; Resnick 1997]. Microworlds provide an excellent context in which to hone and assess students' inquiry skills because they share many features with real apparatus and can make visible important processes that are invisible due to their time and size scale [Gobert 2005a]. The *Science Assistments* group has developed a suite of over 20 microworlds for *Physical, Life, and Earth Science* that are aligned to the *Massachusetts Curricular Frameworks* for Middle school science [Massachusetts Department of Education 2006].

Students are supported through the various phases of inquiry by representational tools, which depict students' data. These representational formats include the *table tool* and the *graph tool*. The table tool supports students by collecting data from trials run by the student; the student can also view this data in order to decide what trials need to be performed. A graph tool was built as a reusable utility that can be included in any microworld. The main purpose of the graph tool is to display any (or several) two-dimensional relations between independent variable(s) and dependent variable(s). The graph tool also includes features to scaffold the student towards the solution.

We also designed several widgets to scaffold students' inquiry; these also support auto-assessment and real time scaffolding. These are important in that they scaffold students in conducting various phases of inquiry, but are also the basis upon which we collect our performance data on students' inquiry skills. Our inquiry widgets were designed in accordance with the learning sciences and science education literature on students' difficulties in conducting inquiry.

For example, research has shown that students have difficulty with inquiry, including choosing the correct variables to work with, forming testable hypotheses, drawing correct conclusions from experiments, and linking hypotheses and data [Chinn and Brewer 1993; Klahr and Dunbar 1988; Kuhn et al. 1995; Kuhn 2005]. They also struggle with translating theoretical variables from their hypotheses into manipulable variables [Richardson 2008], and adequately monitoring what they do [de Jong et al. 2005; de Jong 2006].

Regarding the designing and conducting experiments, a great deal of prior work has shown that students [Chen and Klahr 1999; McElhaney and Linn 2008, 2010; Sao Pedro et al. 2010, 2011], and even scientifically naive adults do not understand the control for variables strategy if not taught explicitly to do so [Kuhn 1991]. These findings emphasize the need for widgets to support students' inquiry processes [Kirschner et al. 2006]. Each of our widgets is described below.

3.1.1 Hypothesis Widget. A key component of scientific inquiry is generating a hypothesis [Kuhn 2005]. In order to support this initial important step of inquiry, the hypothesis widget was developed. This widget provides the capacity for the student to make a hypothesis using a simple

interface. The purpose of the tool is to help students learn how to create hypotheses using independent and dependent variables. Hypotheses are posed as natural-language sentences created through a series of drop-down boxes/multiple-choice widgets. The general structure of the hypothesis is:

When the [*independent variable*] is [*increased/decreased*], the [*dependent variable*] [*increases/decreases/doesn't change*].

Both independent and dependent variables are included in each pull-down box such that sequences of choices permit us an assessment opportunity of whether the student understands (1) what an independent variable is, (2) what a dependent variable is, and (3) the relationship between them. Once the student has generated a hypothesis, they then run trials within the microworld to test their hypothesis; these actions also provide an assessment opportunity as to whether the student tests the hypothesis that they articulated. Further detail on this assessment process and a screen capture of our hypothesis widget (Figure 3) are given later in the paper as we illustrate how our system works.

3.1.2 Interpretation Widget. Just as the hypothesis tool provides the student a structure for creating a hypothesis, the data interpretation tool provides a structure for the student to interpret their data after the experimental trials are completed. Similar to the hypothesis tool, the data interpretation tool presents a way for the student to create statements about the relationship between the independent and dependent variables from their trials using drop-down lists. This also provides a way for the system to auto-score whether students' claims are correct, and whether students can warrant their claims based on data. The second part of this widget provides students with a drop down menu used to warrant their claims using evidence from their experimental trials.

The general structure of the data interpretation widget is:

> When I changed the [*independent variable*] so that it [*increased/decreased*], the [*dependent variable*] [*increased/decreased/didn't change*]. I am basing this on: Data from trial [*trial number from table*] compared to data from trial: [*trial number from table*] this statement [*does support/does not support/is not related to*] my hypothesis. Later in the document we provide a screen capture of this widget (Figure 5).

3.1.3 Communication Widget. Lastly, the students are asked to communicate their findings in open-text format. We incorporate this inquiry step since there is evidence that explaining is a good strategy to get students to reify their understanding [Chi et al. 1989; Chi et al. 1994; Chi 2000]. For communication, we use prompts that are designed to foster the construction of deep explanations; an example of a prompt is:

> "Pretend that you are explaining to a friend the effects of the amount of substance on the boiling point of that substance as if they did not do the experiment."

These prompts are based on findings that students' explanations are deeper when a prompt asks them to explain to someone else [Gobert 2005b].

Currently, we are not autoscoring students' explanations as it is beyond the scope of the project; however, we are reacting to students who write too little, and we are presently developing a detector that will react to those who write gibberish. Briefly, when one of these two situations occurs, Rex our pedagogical agent, intervenes. Our goal here, hence, is not to auto-assess students' communication skills, but to scaffold students who might avoid or trivialize this writing task.

In sum, using our microworlds and the widgets within our environment, students can develop a hypothesis as well as design and conduct an experiment. The resulting data are automatically collected, graphed, and entered into tables so that students can interpret their data and warrant their claims using data; lastly, they communicate their findings.

Next, we present our system and how it relates to ECD; that is, we describe how we used the ECD framework to guide our design and implementation processes for the system and for how to score the resulting sequences of actions. Additionally, we give an example of how the widgets and microworlds work in concert to generate performance data about students' inquiry skills (this in the *task model* section, 4.2).

4. SCIENCE ASSISTMENTS AND ECD

4.1 Student Model

As previously described, the *student model* of the ECD framework consists of aspects of proficiency measured by the assessment (i.e. a collection of variables that represent the skills of a learner) whose values are inferred from students' data [Mislevy et al. 2006]. These are latent skills that we uncover or evaluate by conducting the assessment and those that need to be explicitly modeled to capture learner performance. Put simply, the *student model* asks the question: "What do we want to measure?"

It is worth noting that the ECD definition of *student model* is slightly different from the definition typically used in the ITS and Artificial Intelligence in Education research communities. In particular, these groups do not separate the *evidence model* from the *student model*; the *student model* contains the definition of the latent variables and the ways they and their relationships are measured are combined [cf. Wenger 1987; Bull et al. 1995]. On the other hand, in ECD the *student model* simply defines the latent variables that one intends to measure. The ECD *evidence model* defines how the latent variables will be measured.

For our goals, namely to assess inquiry skills, the variables and components in the *student model* and how we have operationalized them are shown in Figure 1 and described as follows:

4.1.1 Hypothesizing. This entails formulating a testable hypothesis, which requires students to be able to distinguish independent (manipulable) variables from dependent (outcome) variables, and posit some relationship between them. Prior research has shown that students typically have trouble doing so [van Joolingen and de Jong 1991; Njoo and de Jong 1993].
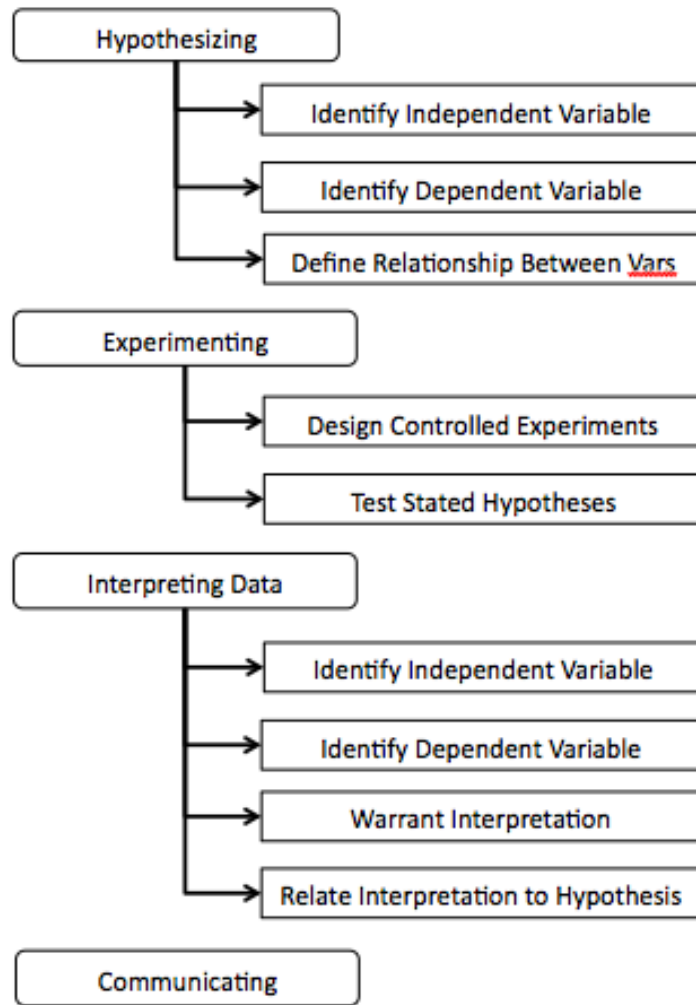
Fig. 1. Skill model with sub-skills for science inquiry.

4.1.2 Designing and Conducting Experiments. This high-level skill entails the process of collecting evidence (data) to enable supporting or refuting hypotheses. We have identified two important sub-skills related to successfully designing and conducting experiments: (1) designing controlled experiments and (2) collecting data to test a stated hypothesis.

Students design controlled experiments when they generate data that can be used to determine the effects of independent variables on outcomes [Sao Pedro et al., 2011]. This entails successful use of the CVS, a strategy stating that one should change only a single variable to be tested, the target variable, while keeping all extraneous variables constant in order to test the effects of that

target variable on an outcome [Chen and Klahr 1999; Kuhn 2005]. It has been noted that students (and adults too) often change too many variables when attempting to test a hypothesis [Tsirgi 1980; Shute and Glaser 1990; Schunn and Anderson 1999; McElhaney and Linn 2008, 2010], and consequently incorrectly draw conclusions from confounded data [Klahr and Dunbar 1988; Kuhn et al. 1992; Schauble et al. 1995].

Students test their stated hypotheses when they generate data with the intent to support or refute an explicitly stated hypothesis [Sao Pedro et al., 2011]. This skill is separated from the designing controlled experiments since students may attempt to test their hypotheses with confounded designs, or may design controlled experiments for a hypothesis not explicitly stated. This skill may be related to students' planning and monitoring [de Jong 2006], particularly if they are testing more than one hypothesis.

4.1.3 Interpreting data. This has been operationalized by our group as correctly interpreting the relationship between the independent and the dependent variable, as per experimental trials collected. Prior research here has shown that students have trouble drawing correct conclusions from experiments, and linking hypotheses and data [Chinn and Brewer 1993; Klahr and Dunbar 1988; Kuhn 2005].

4.1.4 Warranting claims. This has been operationalized by our group as selecting two (or more) trials which either support or refute the hypothesis stated in the hypothesis phase, and demonstrating skill at correctly knowing whether the data support or refute the hypothesis. This is an important skill that relates to epistemological aspects of science (i.e., "knowing how one knows") [Perkins 1986]. Again, students have a great deal of difficulty with this aspect of science inquiry [Chinn and Brewer 1993; Klahr and Dunbar 1988; Kuhn 2005].

4.1.5 Communicating findings. This skill requires that students can express the relationship between cause and effect using evidence and a logical argument. We operationalize this as whether the student included the correct independent variable, dependent variable, and the relationship between them; we also score for the depth of the explanation and whether it includes correct content knowledge. As previously mentioned, this skill is not auto-assessed in our system, as it is beyond the scope of the project.

4.2 Task Model

The *task model* of the ECD framework specifies the key features of the tasks, materials, items, and conditions under which data are collected. This also includes all of the variables that may be required to make sense of learners' actions in the context in which they were performed. Put simply, the *task model* answers the question: "In what situations do we measure it?"

In our system, the microworlds and widgets are used to support inquiry describe the *task model*; we now provide an example of how the microworlds and widgets work together in our *Phase Change* environment (Figures 3 and 4). A typical task provides students with an explicit goal to determine if a particular independent variable (e.g., container size, heat level, substance amount, and cover status) affects various outcomes (e.g., melting point, boiling point, time to melt, and time to boil). Students' inquiry skills (the skills identified in the *student model*) are inferred as they formulate hypotheses, collect data, interpret data, warranting claims with data, and communicating findings about how that variable affected the outcomes within these tasks.

These inquiry processes are supported by guiding students through different inquiry phases: "observe", "hypothesize", "experiment", and "analyze data". Students begin in the "hypothesize" phase and are allowed some flexibility to navigate between phases as shown in Figure 2.
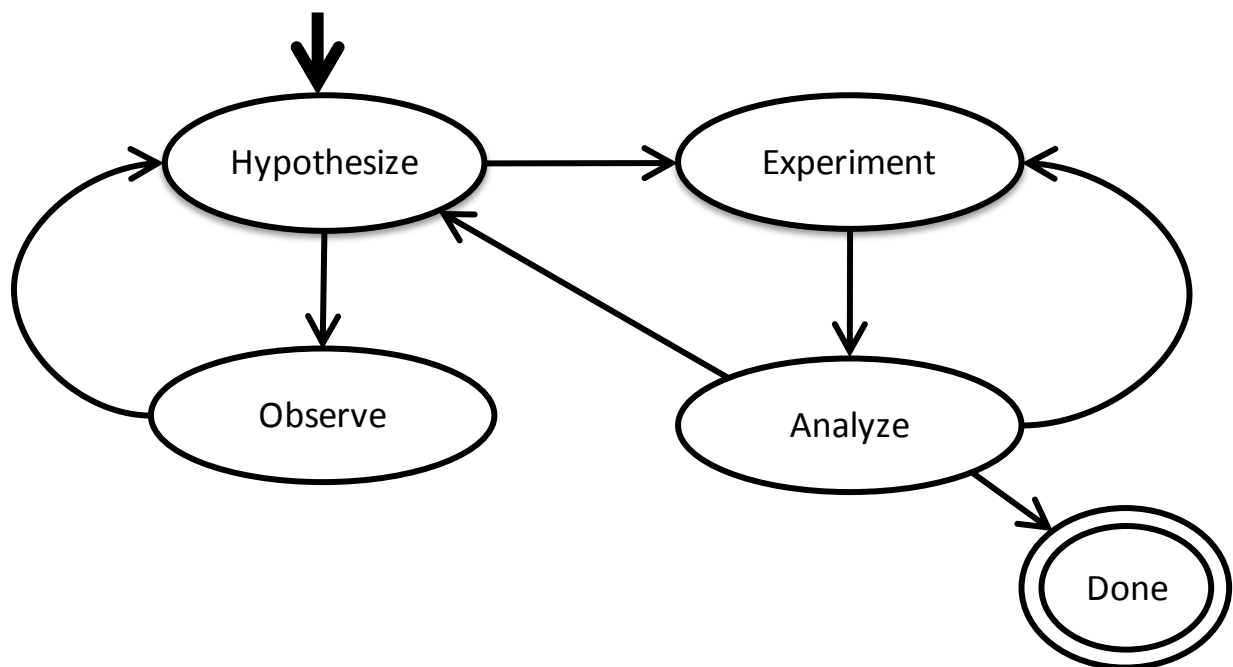

Fig. 2. Paths through inquiry phases.

In the "hypothesize" phase, students use the hypothesis widget (as seen in Figure 3) to generate testable hypotheses. The "observe" phase and "experiment" phase (as seen in Figure 4) are similar. In the "experiment" phase, the student designs and conducts experiments, and has access to two inquiry support tools, a data table summarizing previously run trials, and a hypothesis list. These tools aim to help students plan which experiments to run next.
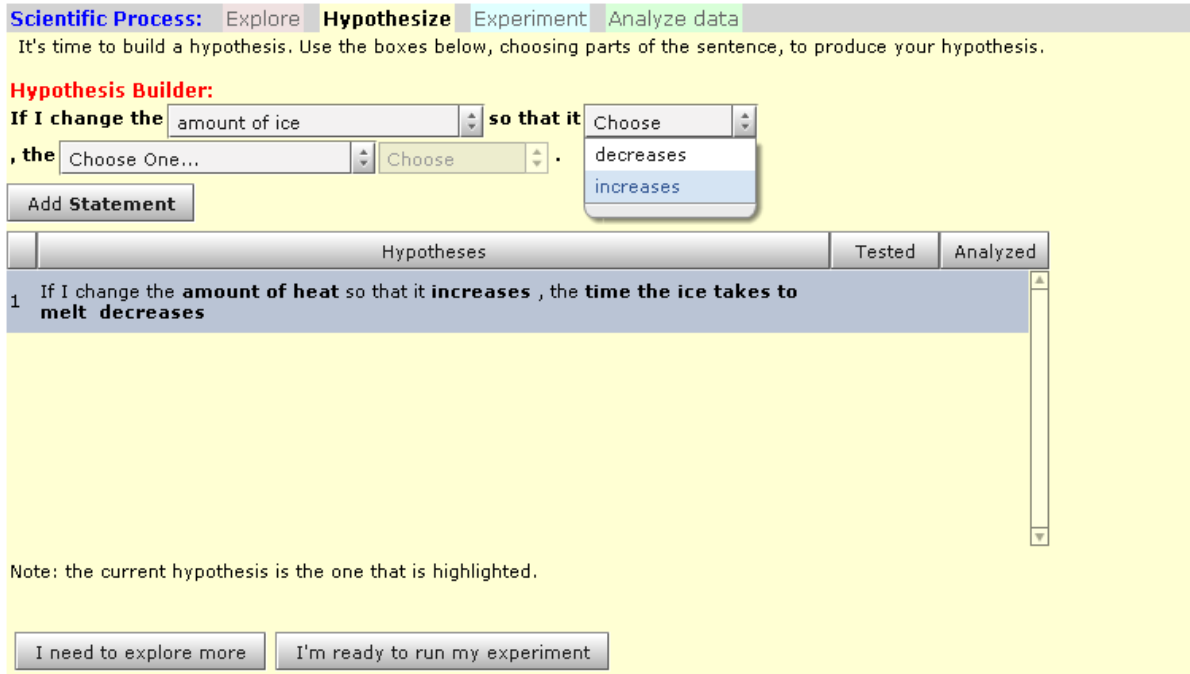

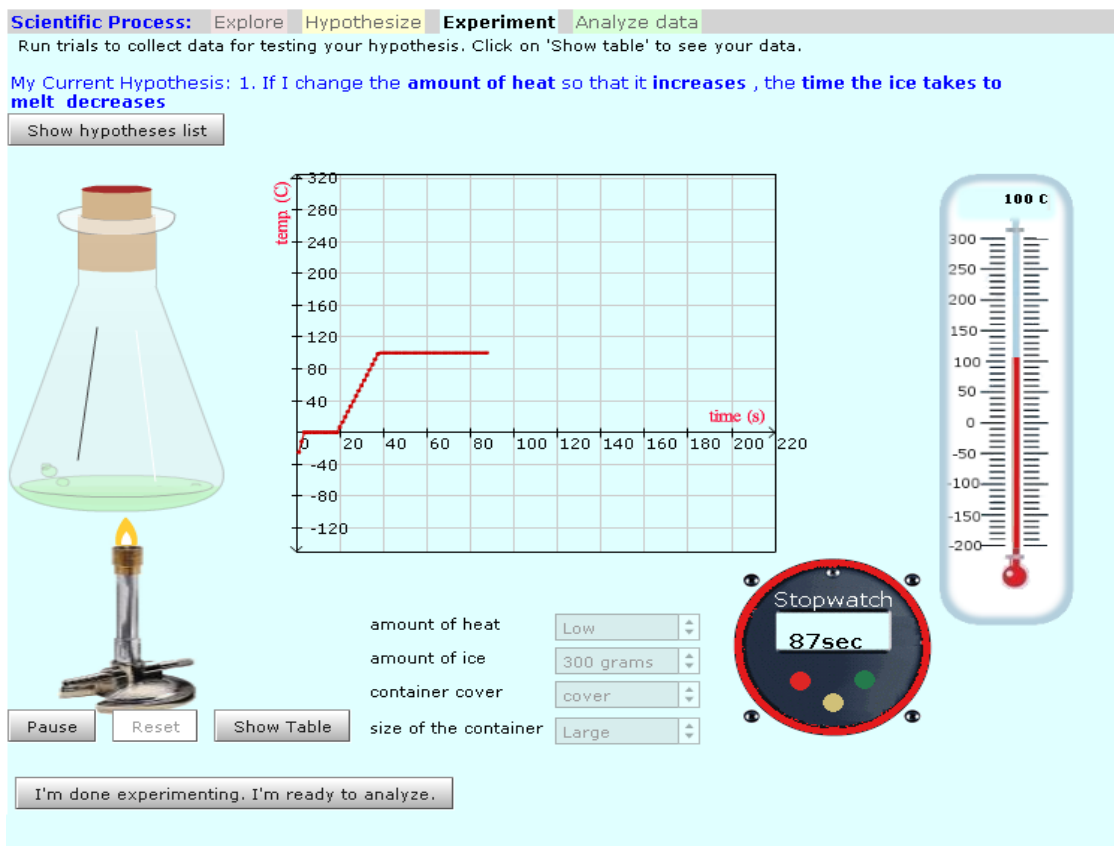Fig. 3. Hypothesizing widget in the *Phase Change* microworld.

Fig. 4. Data collection example for the *Phase Change* microworld.

The "observe" phase, in contrast, hides the inquiry support tools so that students can focus specifically on the simulation. This gives students the opportunity to explore the microworld if they are not yet ready to formulate a hypothesis. Finally, in the "analyze" phase, students are shown the data they collected and use the data analysis tool to construct an interpretation about their data and warrant their claims using data to either support or refute their hypotheses.

As already mentioned, students have some freedom to navigate between inquiry phases (Figure 2) and have flexibility within each phase to conduct many actions. For example, while in the "hypothesize" phase (Figure 3), students could elect to explore the simulation more before formulating any hypotheses by moving to the "observe" phase. Alternatively, they could choose to specify one or more hypotheses like, "If I change the container size so that it increases, the melting point stays the same" before collecting data.

Within the "experiment" phase (Figure 4), students can run as many experiments as they wish to collect data for any one or all of their hypotheses. Within the "analysis" phase, students also have several options (Figure 5). As they construct their claims, students can decide to go back and collect more data or, after constructing claims based on their data, they can decide to create additional hypotheses, thus starting a new inquiry loop. Thus, students can conduct these inquiry activities in various patterns, engaging in inquiry in many different ways.
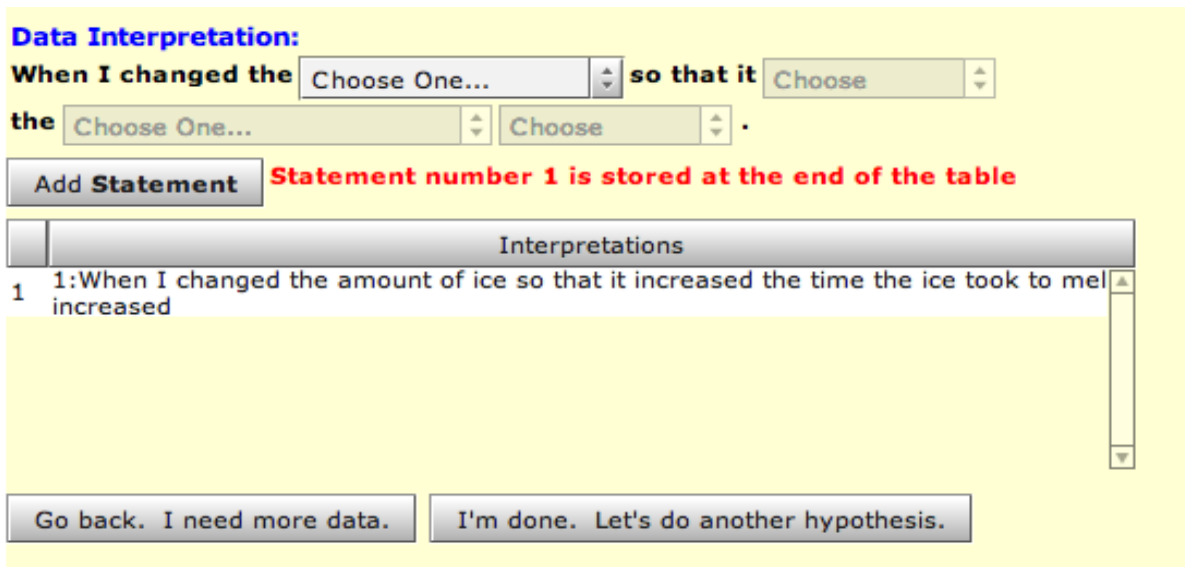

Fig. 5. Interpretation widget in the *Phase Change* microworld.

Since the environment currently does not provide feedback on students' inquiry processes, students can engage in either systematic or haphazard inquiry behavior [Buckley et al. 2006; Buckley et al. 2010]. Specific to the "hypothesize" and "experiment" phases, students acting in a systematic manner [Buckley et al. 2006, 2010] collect data by designing and running controlled experiments that test their hypotheses. They may also use the table tool and hypothesis viewer in order to reflect and plan for additional experiments. These systematic behaviors are representative of the "designing and conducting experiments" skills [NRC 1996, 2011] that we are assessing with machine-learned detectors (described in the *evidence model*).

In contrast, students acting haphazardly in our environment may construct experiments that do not test their hypotheses, not collect enough data to support or refute their hypotheses, design

confounded experiments by not using the control for variables strategy, fail to use the inquiry support tools to analyze their results and monitor their additional experimenting [cf. de Jong 2006], or collect data for the same experimental setup multiple times [Buckley et al. 2006, 2010]. In this paper, we focus on describing how we detect and assess appropriate systematic inquiry behaviors to enable assessment of skills that students are developing, rather than how we detect specific haphazard inquiry behaviors - the latter line of work is described in more detail in Hershkovitz et al. [2011].

4.3 Evidence Model

The *evidence model* is the link between the *task model*, which describes the task, and the *student model*, which describes that which is known about the learner. Specifically, the *evidence model* includes the specifications for the salient features of whatever the learner says, does, or creates in the task situation, as well as the rules for scoring, rating, or otherwise categorizing the salient features of the assessment. Put simply, the *evidence model* answers the question: "How do we measure it?"

Additionally, in the context of performance assessments, Rupp et al. [this issue] distinguish between *product data* and *process data*. Product data are essentially students' finalized, tangible work products, whereas process data come from learners' interactions with other learners or interactions within the environment. In our microworlds, product data are akin to students' stated hypotheses and results of the analyses of their data. Process data, on the other hand, arise from students' entire stream of data-collection activities.

Since we have two kinds of data, we use two different methodologies for assessing students. Our product data is assessed using knowledge-engineered rules whereas our process data is assessed using more complex behavior models derived using techniques from EDM [cf. Baker and Yacef 2009; Romero and Ventura 2010].

4.3.1 Assessing Product Data: Hypothesizing and Analyzing. The "constructing hypotheses" skill - not to be confused with the "testing hypotheses" skill - and the data interpretation skill are measured by assessing students' data; we refer to these as work products. These work products, in particular, are the hypotheses students generate using the hypothesis widget, and the analyses they construct about their data using the analysis widget. These products lend themselves to

evidence identification processes using knowledge-engineered rules because they are relatively well-defined skills (see Figure 1).

For example, when a student generates an analysis of the data that they collect, our system uses knowledge-engineered rules to "check" that the analysis has the following components: did they identify the correct independent variable and dependent variable, did they specify the correct relationship between them, and lastly, did the appropriate trials to support their analysis.

Currently, we have implemented the rules to automatically score students' hypotheses and data interpretations. We generate binary assessments for each hypothesizing sub-skill each time a student formulates a hypothesis, and binary assessments for each interpreting data sub-skill each time a student constructs an analysis and warrants their claims. However, to date, we have not yet developed models to *aggregate* these assessments across inquiry activities and domains to determine the degree to which students have mastered these skills. In the future, we plan on doing so by employing *Bayesian Knowledge-Tracing* (BKT) [Corbett and Anderson 1995].

4.3.2 Assessing Process Data: Experimenting. To assess and track the two skills "designing controlled experiments" and "testing stated hypotheses", we use the following tools for the associated process data. First, we apply our validated, machine-learned detectors (i.e., models) of behaviors associated with these skills to classify students' actions as they collect data. Essentially, these models encapsulate evidence identification rules [Mislevy et al. 2006] whose resulting classifications are the assessments of skill demonstrated in a practice opportunity.

As students engage in inquiry over several activities, the detectors are applied each time the student collects data. We then feed the detectors' classifications into BKT models associated with each skill to keep track of students' proficiency over time. To date, we have developed and validated such models for our phase change physical science microworld [Sao Pedro et al. 2010; Montalvo et al. 2010; Sao Pedro et al. 2011]. Below, we briefly describe how they were developed and validated for that microworld.

4.3.2.1 Development of Behavior Models. To generate behavior models of each skill, we utilized *text replay tagging* of log files [Sao Pedro et al. 2010; Montalvo et al. 2010; Sao Pedro et al. 2011], which is an extension of the text replay approach developed in Baker et al. [2006]. Similar to a video replay or screen replay, a text replay uses a pre-specified chunk of student

actions presented in text format that includes information such as an associated time stamp, action type, widget type, and exact input.

Text replay tagging differs from other types of text replay in one key way. Text replays allow for the classification of a replay clip as a single category out of a set of categories. Text replay tagging, in contrast, allows multiple tags to be associated with one clip. For example, within the domain of science inquiry, a clip may be tagged as involving designing controlled experiments, involving testing stated hypotheses, both, or neither.

In text replay tagging, human coders first hand-coded clips gleaned from students' engagements with the phase change microworld activities. *Clips*, contiguous sequences of low-level student actions, were labeled using a set of behavior tags (i.e., "designed controlled experiments" and/or "tested stated hypotheses"). A clip (shown in Figure 6) contained all interface interactions within the hypothesizing and experimenting phases of inquiry (see Figure 2) since these phases are relevant to labeling data collection behaviors. Then, a feature set was distilled to summarize clips and, with the tags, were used to construct machine-learned detectors for each behavior.
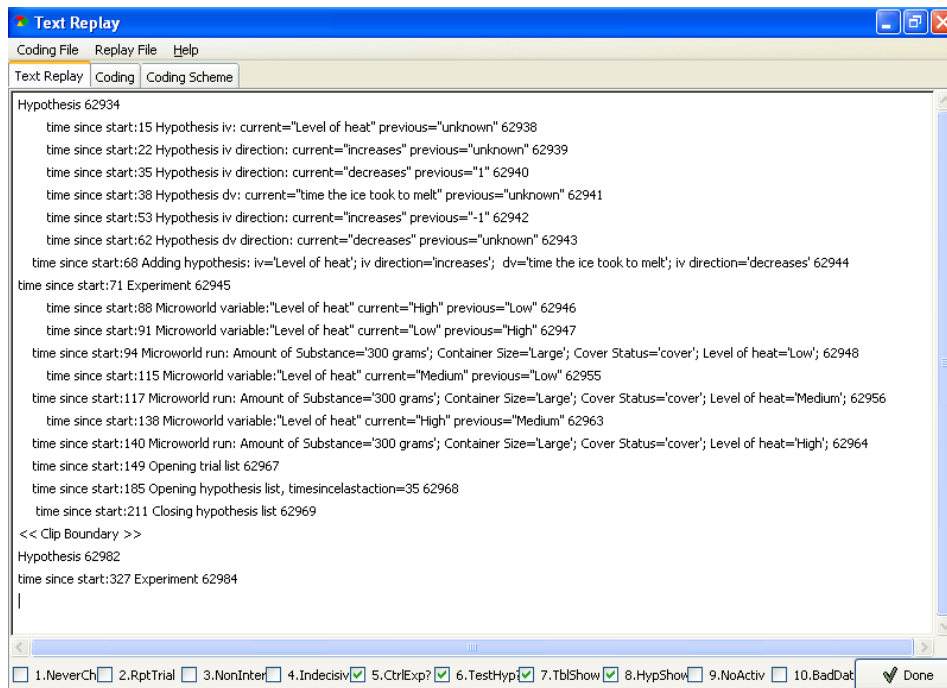

Fig. 6. An example clip labeled by a human coder. This clip was tagged as involving designing controlled experiments and testing stated hypotheses, in addition to other behaviors.

Following this methodology, detectors for the "designing controlled experiments" behavior and "testing stated hypotheses" behavior were constructed and validated. To be brief, we present here only high-level details and a summary of the results from Sao Pedro et al. [2011]. A full description of how these models were generated from fine-grained student logs appears in that publication.

Detectors were constructed using data gleaned from 148 middle school students' interactions with the phase change microworld. Approximately 570 clips were hand-coded by two coders to generate a set of training instances from which to build detectors. Detectors were built by removing correlated features and using J48 decision trees with automated pruning in order to control for over-fitting. They were validated using six-fold student-level cross validation and had their predictive goodness assessed using two metrics, A'[1] [Hanley and McNeil 1982] and *Cohen's Kappa* (κ)[2] [Cohen 1960]. An illustration and interpretation of the results decision trees for each skill appear in Sao Pedro et al. [2011].

Overall, the results for classifying student behavior with the detectors were very promising. The detectors could distinguish a set of trials in which a student designed controlled experiments from a set of trials in which students did not design controlled experiments 85% of the time. They could also distinguish a set of trials in which a student tested their stated hypotheses from a set of trials in which they did not 85% of the time. Furthermore, the associated Kappa values, ranging from .40 to .47, indicated that each of these detectors was better than chance. The performance of these detectors, as measured by A' and Kappa, is comparable to detectors of gaming the system refined over several years [e.g., Baker and de Carvalho 2008; Baker et al. 2010]. Therefore, these detectors can be used to automatically classify students' data collection behavior.

---

[1] A' is the probability that if the detector is comparing two clips, one involving the category of interest (designing controlled experiments, for instance) and one not involving that category, it will correctly identify which clip is which. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly.

[2] Cohen's Kappa assesses whether the detector is better than chance at identifying the correct action sequences as involving the category of interest. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly.

4.3.2.2. Development of Skill Proficiency Estimates. The behavior detectors have the capability to distinguish students who are engaging in systematic data collection behavior from those who are not within a clip. Clips are the granularity at which these behaviors are exhibited, and as such, can also be viewed as a practice opportunity. To amalgamate students' performances over practice opportunities (i.e., clips) and produce proficiency skill estimates, we used BKT due to its success at estimating skill in other learning environments that provide explicit learning support [e.g. Corbett and Anderson 1995; Koedinger and Corbett 2006; Baker et al. 2008a; Ritter et al. 2009; Feng et al. 2009; Baker et al. 2010; Pardos et al. 2010].

Briefly, a BKT model is a two-state hidden Markov model that estimates the probability a student possesses latent skill (L_n) after *n* observable practice opportunities. In our case, observable student performance is demonstration of data collection skill within a clip, as assessed by the detectors. BKT models are characterized by four parameters, G, S, L_0, and T, used in part to compute latent skill (L_n). The *guess* parameter (G) is the probability the student will demonstrate the skill despite not knowing it. Conversely, the *slip* parameter (S) is the probability the student will not demonstrate the skill even though they know it. Finally, L_0 is the initial probability of knowing the skill before any practice and T is the probability of learning the skill between practice attempts. Within the BKT framework, these four parameters are assumed to be the same for all students and knowledge of the skill is assumed to be binary, i.e. either the student knows the skill, or they do not.

Using these four parameters, the probability that a student knows the skill, $P(L_n)$ and the estimate that a student will demonstrate that skill in their next practice opportunity $P(Demonstrate\_Skill_n)$ can be computed. The equations for these calculations are:

$$P(L_{n-1}|Obs_n) = \begin{cases} \dfrac{P(L_{n-1}) * (1-S)}{P(L_{n-1}) * (1-S) + \left(1 - P(L_{n-1})\right) * G}, Demonstrated\_Skill_n \\[3em] \dfrac{P(L_{n-1}) * S}{P(L_{n-1}) * S + \left(1 - P(L_{n-1})\right) * (1-G)}, {\sim}Demonstrated\_Skill_n \end{cases}$$

$$P(L_n) = P(L_{n-1}|Obs_n) + \left(\left(1 - P(L_{n-1}|Clip_n)\right) * T\right)$$

$$P(Demonstrate\_Skill_n) = P(L_{n-1}) * (1-S) + (1 - P(L_{n-1})) * T$$

Note that *P(Demonstrate_Skill_n)* is an a-priori estimate of demonstrating skill since it depends on the prior estimate of knowing the skill, $P(L_{n-1})$. An example application of these equations for a students' performance profile associated with the designing controlled experiments skill is shown in Table I. In this example, the student generated 9 clips over all activities (and thus engaged in 9 data collections). Each clip was labeled as demonstrating skill or not using the designing controlled experiments detector. From there, estimates for $P(L_n)$ and *P(Demonstrate_Skill_n)* can be found by applying the equations above.

Table I. Example Student Practice Profile with BKT Estimates

Designing Ctrl'd Exp's Practice Opportunities	1	2	3	4	5	6	7	8	9	Final
$P(L_{n-1})$	0.077	0.387	0.819	0.97	0.795	0.965	0.769	0.959	0.994	0.999
P(Demonstrate_Skill)	0.191	0.429	0.761	0.877	0.742	0.873	0.723	0.869	0.895	
Observable: Demonstrated Skill?	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	

BKT Model: $\{L_0 = .077, G = .132, S = .100, T = .038\}$

Note. This student engaged in 9 data collection activities, and their final estimate of knowing this skill is $P(L_n) = .999$.

In Sao Pedro et al. [2011], we used *brute force search* [cf. Baker et al. 2011] to find the best fitting parameter estimates (values for G, S, L_0, and T) given data for the phase change microworld.

We then validated the goodness of the resulting BKT models for each skill in two ways. First, we predicted performance within the environment, *P(Demonstrate_Skill_n)*, providing a measure of the internal reliability. On this metric, it was found that BKT could estimate skill at each practice opportunity acceptably (A' = .74 for designing controlled experiments and A' = .79 for testing stated hypotheses). Second, we used the BKT probabilistic estimates of knowing the skills (L_n) to predict performance on transfer tasks requiring inquiry skill, providing a measure of external validity. Overall, each model of authentic inquiry skill was significantly, albeit modestly, correlated to its corresponding transfer test (i.e., with the standardized-test style

questions on hypotheses). This provided some external validation of the skill estimates derived from performance within the phase change environment. Thus, the BKT models appeared to be a valid way to amalgamate evidence across activities in order to estimate students' skills.

5. DISCUSSION

In response to calls for inquiry assessments, we designed our educational environment, *Science Assistments*, which assesses students' inquiry skills. The ECD framework [Mislevy and Haertel 2006] was used to help our initial conceptualizations of the *student model*, the *task model*, and the *evidence model*, as well as to help guide our data analyses.

Specifically, for us, the *student model* included the specifications of the skills and sub-skills of inquiry; the *task model* included the specifications of the tasks, items, conditions, and forms completed by the student in our environment; and the *evidence model* included the data, in our case extracted from the log files of students' inquiry regarding whether students area testing their articulated hypotheses or are designing controlled experiments. We have, in this paper, provided an overview about how inquiry skills other than those involved in designing and conducting experiments can be conceptualized and analyzed with ECD as well.

In designing our learning environment / microworlds, we balanced the need for creating rich, authentic inquiry tasks and the goal of auto-scoring students' inquiry skills. This was a complicated task and many design trade-offs that are currently being discussed in the learning science literature influenced our design [Kirschner et al. 2006; Hmelo-Silver et al. 2007]. Specifically, we addressed the optimal degree of desired open-endedness versus degree of guidance so that students' inquiry skills could be honed, as described by knowledge ontologies of science content, inquiry skills, and understanding of the nature of science [Perkins 1986]. As a consequence of our design approach and decisions, our environment is fairly flexible in terms of how students can approach an inquiry task.

This design also made our auto-scoring of inquiry skills more complicated in some ways. For example, in the hypothesis phase, students can either list all their hypotheses at once or generate and test them one at a time. Additionally, as previously discussed, students also can choose to return to an earlier phase of inquiry or choose to explore the phenomena a second time in order to specify a hypothesis.

We have evidence that students do not conduct inquiry in a lock-step fashion. For example, Bachmann et al. [2010, 2011; Bachmann 2012] showed that many students, when given an orienting goal for their inquiry, conducted virtually all their trials in the "explore" phase of the inquiry process, and in doing so, arrived at the correct solution before entering the experimenting phase. Additionally, Gobert et al. [in preparation] analyzed data in which students, when asked to warrant their claims with data, returned to the experiment phase and were much more systematic in their data collection the second time around. This was evidenced by more systematic use of the control for variables strategy and targeting the proper independent variable in order to test their articulated hypothesis or the one specified by the orienting question. Data such as these serve to illustrate the flexibility of our system for students' inquiry, but also serves to illustrate the complexity of assessing students' inquiry skills within such an environment.

As previously stated, we used the NSES inquiry strands [NRC 1996] to inform the design of our environment and our widgets. We designed the widgets, in part, by operationalizing the inquiry strands into sub-skills. Since there was a great deal of prior literature on students' difficulties in conducting inquiry, a rational analysis of each skill allowed us to operationalize these into sub-skills. We also relied on think-aloud protocols collected with individual students while they conducted inquiry in order to ascertain the aspects of each skill with which students were having difficulty [Richardson 2008; Gobert et al. 2008].

Our widgets, originally designed to support students' inquiry, provide the affordance of generating log data that enabled us to assess students' work products and experimentation processes. We developed knowledge engineered rules by hand to assess students' work products, namely, their hypotheses and data interpretations. Knowledge engineering was chosen since "correctness" for these skills was more well-defined than students' data collection processes, on the other hand, which were assessed using machine-learned models to account for the ill-defined nature of what it means to properly collect data. These assessments in total then can be amalgamated to form estimates of whether students know each inquiry skill over several inquiry activities of the same type. In the present work, we showed how we did this for data collection skills using BKT.

The previously described assessment and skill estimation models can be leveraged for a variety of purposes. First, they can be used to generate assessment reports for teachers that can be viewed / aggregated in various ways, including the class level and the individual student level. Such formative assessment feedback can enable teachers to make pedagogical decisions in real time, knowing which students and which inquiry sub-skills to focus on. Thus, our widgets support a first critical step towards developing the means to auto-assess students as they conducted inquiry; this is an advantage over other approaches in that we can track these skills in a fine-grained way. Lastly, these data also can be used to provide real time scaffolding to students on their inquiry as well, via our pedagogical agent Rex [Gobert and Baker 2012].

Our 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 et al. 1999], (2) the capacity to assess inquiry skills in the context in which they are developing [Mislevy et al. 2003], and (3) providing metrics that reflect the validity of these authentic measures of inquiry skills [Williamson et al. 2006].

6. CONCLUSIONS & NEXT STEPS

In this paper we presented a description of our approach within the context of the ECD framework. We utilized log files of students' data collected in real time and EDM techniques which we applied to these log files to characterize students in terms of their proficiency in the area of science inquiry skills, specifically testing hypotheses as well as designing and conducting experiments.

EDM techniques can capture students' inquiry performances and auto-score them in a manner that handles their complexity. For example, in the behavior models we developed, we are able to differentiate whether a student knows the control for variables strategy even when the student chooses not to conduct sequential trials. Other projects, which rely principally on rule-based methods for scoring data, run the risk of either being too stringent. For example, they may count only sequential trials in which the target variable is changed as evidence of CVS skill [McElhaney and Linn 2008, 2010], or may be overly "generous" by comparing all trials to all other trials in order to determine whether the student has demonstrated the CVS skill [Gobert and

Koedinger 2011] potentially scoring coincidental matches between trials as evidence of CVS skill.

Furthermore, conducting text replays and machine learning over multiple experimental trials per student as described in the present paper allowed us to automatically evaluate those who know CVS versus those who do not by using only a small number of trials. That is, when a student is conducting multiple trials that are non-sequential in their use of CVS, we can evaluate whether the student has demonstrated the skill.

A key step in the development of detectors as general measurements of constructs such as those discussed earlier is validating the generalizability of the detectors. A methodological template for our planned process can be found in Baker et al. [2008a], which studied the generalization of detectors of gaming the system. There are at least three essential forms of generalizability for a detector to be fully validated for *Science Assistments*, (1) student-level, (2) classroom-level, and (3) microworld-level.

We have already validated that detectors of student inquiry behaviors in *Science Assistments* can generalize between students. What remains to be done is the validation of the detectors at the classroom-level and microworld-level. To do this, it will be necessary to collect data for additional classrooms and microworlds, develop training labels using text replays, and then conduct cross-validation at the both classroom and the microworld level. Our past experience in developing detectors of gaming the system [Baker et al. 2008a] suggests that conducting microworld or unit-level cross-validation is more effective when there is a larger sample of units to cross-validate between -- excellent performance was achieved in that case when cross-validating across four units.

To this end, we plan to collect and label data for three additional microworlds beyond the phase change unit already studied. Once validation of our detectors is developed for three additional physical science microworlds, we can in turn generalize these to the other science domains addressed in *Science Assistments*, namely *Life Science* and *Earth Science* topics. Lastly, we will continue our work in the development of other detectors to auto-assess students' inquiry skills for the other skills in the *National Science Education Standards* [NRC 1996], giving us a full suite of detectors with which to auto-assess and auto-scaffold students' science inquiry across all middle school topics.

REFERENCES

AGRAWAL, R., AND SRIKANT, R. 1994. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20ᵗʰ VLDB Conference*, Santiago, Chile, 487-499.

ALMOND, R.G., WILLIAMSON, D.M., MISLEVY, R.J., AND YAN, D. In press. *Bayes nets in educational assessment*. Springer, New York, NY.

ALONZO, A., AND ASCHBACHER, P.R. 2004. *Value Added? Long assessment of students' scientific inquiry skills*. Presented at *the Annual Meeting of the American Educational Research Association*, San Diego, CA.

AMERSHI, S., AND CONATI, C. 2009. Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining, 1* (1), 71-81.

ANDERSON, J.R., AND LEBIERE, C. 1998. *The atomic components of thought.* Erlbaum, Mahwah, NJ.

BACHMANN, M. 2012. Biology Microworld to Assess Students' Content Knowledge and Inquiry Skills and Leveraging Student modeling to Prescribe Design Features for Scaffolding Learning. Unpublished Master's thesis. Worcester Polytechnic Institute, Worcester, MA.

BACHMANN, M., GOBERT, J.D., AND BECK, J. 2010. Tracking Students' Inquiry Paths through Student Transition Analysis. In *Proceedings of the 3rd International Conference on Educational Data Mining*, 269-270.

BACHMANN, M., GOBERT, J., AND BECK, J. 2011. Do Differences in Student's Exploration Behavior lead to differences in Domain Learning or Inquiry Skills? Presented at *the Annual Meeting of the American Educational Research Association*, New Orleans, LA.

BAKER, R.S.J.D., CORBETT, A.T., AND ALEVEN, V. 2008a. More Accurate Student modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 406-415.

BAKER, R.S.J.D., CORBETT, A.T., GOWDA, S.M., WAGNER, A.Z., MACLAREN, B.M., KAUFFMAN, L.R., MITCHELL, A.P., AND GIGUERE, S. 2010. Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 52-63.

BAKER, R.S., CORBETT, A.T., ROLL, I., AND KOEDINGER, K.R. 2008b. Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction , 18* (3), 287-314.

BAKER, R., CORBETT, A., AND WAGNER, A. 2006. Human Classification of Low-Fidelity Replays of Student Actions. In *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, 29-36.

BAKER, R., AND DE CARVALHO, A. 2008. Labeling Student Behavior Faster and More Precisely with Text Replays. In *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, R.S. BAKER, T. BARNES, AND J.E. BECK, Eds. Montreal, Quebec, Canada, 38-47.

BAKER, R.S., MITROVIC, A., AND MATHEWS, M. 2010. Detecting Gaming the System in Constraint-Based Tutors. In *Proceedings of the 18th Annual Conference on User Modeling,*

Adaptation, and Personalization, UMAP 2010. LNCS 6075, P. DE BRA, P. KOBSA, AND D. CHIN, Eds. Springer-Verlag, Big Island of Hawaii, HI, 267-278.

BAKER, R.S.J.D., PARDOS, Z., GOWDA, S., NOORAEI, B., AND HEFFERNAN, N. 2011. Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems. *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization*, 13-24.

BAKER, R., AND YACEF, K. 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1* (1), 3-17.

BAUM, L.E., AND PETRIE, T. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics, 37* (6), 1554-1563.

BAXTER, G., AND SHAVELSON, R. 1994. Science performance assessments: benchmarks and surrogates. *International Journal of Education Research, 21* (3) , 279-298.

BERNARDINI, A., AND CONATI, C. 2010. Discovering and Recognizing Student Interaction Patterns in Exploratory Learning Environments. In *Proceedings of the 10th International Conference of Intelligent Tutoring Systems, ITS 2010, Part 1*, V. ALEVEN, J. KAY, AND J. MOSTOW, Eds. Springer-Verlag, Berlin Heidelberg, 125-134.

BLACK, P. 1999. *Testing: Friend or Foe? Theory and Practice of Assessment and Testing*. Falmer Press, New York, NY.

BRYSON, A.E., AND HO, Y.-C. 1969. *Applied Optimal Control*. Blaisdell, New York.

BUCKLEY, B. C., GOBERT, J.D., AND HORWITZ, P. 2006. Using log files to track students' model-based inquiry. In *Proceedings of the 7th International Conference on Learning Sciences, ICLS 2006*, Erlbaum, Bloomington, IN, 57-63.

BUCKLEY, B., GOBERT, J., HORWITZ, P., AND O'DWYER, L. 2010. Looking Inside the Black Box: Assessments and Decision-making in BioLogica. *International Journal of Learning Technology, 5* (2), 166-190.

BULL, S., BRNA, P., AND PAIN, H. 1995. Extending the scope of the *student model*. *User Modeling and User-Adapted Interaction, 5* (1), 45-65.

CHAMPAGNE, A., BERGIN, K., BYBEE, R., DUSCHL, R., AND GALLAGHER, J. 2004. NAEP 2009 science framework development: Issues and recommendations. Paper commissioned by the National Assessment Governing Board, Washington, DC.

CHEN, Z., AND KLAHR, D. 1999. All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development, 70* (5), 1098-1120.

CHI, M. 2000. Self-explaining Expository Texts: The Dual Process of Generating Inferences and Repairing Mental Models. In *Advances in Instructional Psychology*, R. GLASER, Ed. Lawrence Erlbaum Associates, Inc., Mahweh, NJ, 161-238.

CHI, M., BASSOK, M., LEWIS, M.W., REIMANN, P., AND GLASER, R. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145-182.

CHI, M., DELEEUW, N., CHIU, M., AND LAVANCHER, C. 1994. Eliciting Self-Explanations Improves Understanding. *Cognitive Science, 18*, 439-477.

CHINN, C.A., AND BREWER, W.F. 1993. The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*, 1-49.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological*

Measurement, 20 (1), 37-46.

CORBETT, A., AND ANDERSON, J. 1995. Knowledge-Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

CROCKER, L., AND ALGINA, J. 2006. *Introduction to Classical and Modern Test Theory*. Cengage Learning, Independency, KY.

DE AYALA, R.J. 2009. *The Theory and Practice of Item Response Theory*. The Guilford Press, New York, NY.

DE JONG, T. 2006. Computer simulations - Technological advances in inquiry learning. *Science, 312*, 532-533.

DE JONG, T., BEISHUIZENM, J., HULSHOF, C., PRINS, F., VAN RIJN, H., VAN SOMEREN, M., ET AL. 2005. Determinants of Discovery Learning in a Complex Simulation Learning Environment. In *Cognition, Education and Communication Technology,* P. GARDENFORS, AND P. JOHANSSON, Eds. Lawrence Erlbaum Associates, Mahwah, NJ, 257-283.

DE JONG, T., VAN JOOLINGEN, W., GIEMZA, A., GIRAULT, I., HOPPE, U., KINDERMANN, J., ET AL. 2010. Learning by creating and exchanging objects: The SCY experience. *British Journal of Educational Technology, 41* (6), 909-921.

FADEL, C., HONEY, M., AND PASNICK, S. 2007. Assessment in the Age of Innovation. *Education Week, 26* (38), 34-40.

FENG, M., HEFFERNAN, N.T., AND KOEDINGER, K.R. 2009. Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI), 19* (3), 243-266.

GHAZARIAN, A., AND NOORHOSSEINI, S.M. 2010. Automatic Detection of Users' Skill Levels Using High-Frequency User Interface Events. *User Modeling and User-Adapted Interaction, 20* (2), 109-146.

GLASER, R., SCHAUBLE, L., RAGHAVAN, K., AND ZEITZ, C. 1992. Scientific Reasoning Across Different Domains. In *Computer-based Learning Environments and Problem-Solving*, E. DECORTE, M. LINN, H. MANDL, AND L. VERSCHAFFEL, Eds. Springer-Verlag, Heidelberg, Germany, 345-371.

GOBERT, J. 2005a. Leveraging Technology and Cognitive Theory on Visualization to Promote Students' Science Learning and Literacy. In *Visualization in Science Education*, J. GILBERT, Ed. Springer-Verlag Publishers, Dordrecht, The Netherlands, 73-90.

GOBERT, J. 2005b. 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., AND BAKER, R. 2010. *Empirical Research: Emerging Research: Using Automated Detectors to Examine the Relationships Between Learner Attributes and Behaviors During Inquiry in Science*. Proposal Awarded July 1, 2010 by the National Science Foundation.

GOBERT, J., AND BAKER, R. 2012. *The Development of an Intelligent Pedagogical Agent for Physical Science Inquiry Driven by Educational Data Mining*. Proposal (R305A120778) awarded May, 2012 by the U.S. Dept. of Education.

GOBERT, J., HEFFERNAN, N., KOEDINGER, K., AND BECK, J. 2008. *ASSISTments Meets Science Learning* (AMSL). Proposal (R305A090170) funded February 1, 2009 by the U.S. Dept. of Education.

GOBERT, J., HEFFERNAN, N., RUIZ, C., AND RYUNG, K. 2007. *AMI: ASSISTments Meets Inquiry*. Proposal NSF-DRL# 0733286 funded by the National Science Foundation.

GOBERT, J., AND KOEDINGER, K. 2011. *Using model-tracing to conduct performance assessment of students' inquiry skills within a Microworld.* Presented at the Society for Research on Educational Effectiveness, Washington, D.C., September 8-10.

GOBERT, J., RAZIUDDIN, J., AND MONTALVO, O. In prep. Warranting claims as an epistemological driver. Manuscript in preparation.

GOTWALS, A., AND SONGER, N. 2006. Measuring Students' Scientific Content and Inquiry Reasoning. In *Proceedings of the 7th International Conference of the Learning Sciences, ICLS 2006*, S. BARAB, K. HAY, AND D. HICKEY, Eds. Lawrence Erlbuam Associates, Bloomington, IN, 196-202.

HAMBLETON R., AND JONES, R. 1993. Comparison of Classical Test Theory and Item Response Theory ad their applications to test development. *Educational Measurement: Issues & Practice 12* (3), 38-47.

HANLEY, J., AND MCNEIL, B. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology, 143*, 29-36.

HARRISON, A.M., AND SCHUNN, C.D. 2004. The transfer of logically general scientific reasoning skills. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, K. FORBUS, D. GENTNER, AND T. REGIER, Eds. Erlbaum, Mahwah, NJ, 541-546.

HEFFERNAN, N., TURNER, T., LOURENCO, A., MACASEK, M., NUZZO-JONES, G., AND KOEDINGER, K. 2006. The ASSISTment builder: Towards an analysis of cost effectiveness of ITS creation. In *Proceedings of the 19th International FLAIRS Conference*, Melbourne Beach, FL, 515-520.

HERSHKOVITZ, A., WIXON, M., BAKER, R.S.J.D., GOBERT, J., AND SAO PEDRO, M. 2011. Carelessness and Goal Orientation in a Science Microworld. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 462-465.

HMELO-SILVER, C.E., DUNCAN, R.G., AND CHINN, C.A. 2007. Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Krischner, Sweller, and Clark (2006). *Educational Psychologist, 42* (2), 99-107.

KIRSCHNER, P.A., SWELLER, J., AND CLARK, R.E. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discover, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41* (2), 75-86.

KLAHR, D., AND DUNBAR, K. 1988. Dual search space during scientific reasoning. *Cognitive Science, 12* (1), 1-48.

KOEDINGER, K., ANDERSON, J., HADLEY, W., AND MARK, M. 1997. Intelligent Tutoring Goes to School in the Big City. *International Journal of Arificial Intelligence in Education, 8*, 30-43.

KOEDINGER, K., AND CORBETT, A. 2006. Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In *The Cambridge Handbook of the Learning Sciences*, R. SAWYER, Ed. Cambridge University Press, New York, NY, 61-77.

KOEDINGER, K., SUTHERS, D., AND FORBUS, K. 1998. Component-Based Construction of a Science Learning Space. *International Journal of Artificial Intelligence in Education (IJAIED), 10*, 292-313.

KRAJCIK, J., BLUMENFELD, P., MARX, R., BASS, K., FREDRICKS, J., AND SOLOWAY, E. 1998. Inquiry in project-based science classrooms: Initial attempts by middle school students. *Journal of the Learning Sciences, 7,* 313-350.

KUHN, D. 1991. *The skills of argument*. Cambridge Press, Cambridge, MA.

KUHN, D. 2005. *Education for thinking.* Harvard University Press, Cambridge, MA.

KUHN, D., GARCIA-MILA, M., ZOHAR, A., AND ANDERSEN, C. 1995. Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60* (4, Serial No. 245).

KUHN, D., SCHAUBLE, L., AND GARCIA-MILA, M. 1992. Cross-domain development of scientific reasoning, *Cognition and Instruction, 9* (4), 285-327.

LIU, B., HSU, W., AND MA, Y. 1998. Integrating Classification and Association Rule Mining. In *Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, 80-86.

MARTIN, J., AND VANLEHN, K. 1995. Student assessment using Bayesian nets. *International Journal of Human-Computer Studies, 42* (6), 575-591.

MASSACHUSETTS DEPARTMENT OF EDUCATION. 2006. *Massachusetts Science and Technology/Engineering Curriculum Framework.* Massachusetts Department of Education, Malden, MA.

MCELHANEY, K., AND LINN, M. 2008. Impacts of Students' Experimentation Using a Dynamic Visualization on their Understanding of Motion. In *Proceedings of the 8th International Conference of the Learning Sciences, ICLS 2008, Volume 2*, International Society of the Learning Sciences, Inc., Utrecht, The Netherlands, 51-58.

MCELHANEY, K., AND LINN, M. 2010. Helping Students Make Controlled Experiments More Informative. In *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010) - Volume 1, Full Papers*, K. GOMEZ, L. LYONS, AND J. RADINSKY, Eds. International Society of the Learning Sciences, Chicago, IL, 786-793.

MCNEILL, K.L., AND KRAJCIK, J. 2007. Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In *Thinking with data*, M. LOVETT AND P. SHAH, Eds. Taylor & Francis Group, LLC, New York, NY, 233-265.

MESSICK, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13-23.

MISLEVY, R.J., BEHRENS, J.T., DICERBO, K.E., AND LEVY, R., This issue. Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining.*

MISLEVY, R., CHUDOWSKY, N., DRANEY, K., FRIED, R., GAFFNEY, T., AND HAERTEL, G. 2003. *Design Patterns for Assessing Science Inquiry*, SRI International, Menlo Park, CA.

MISLEVY, R.J., AND HAERTEL, G.D. 2006. Implications of ECD for educational testing. *Educational Measurement: Issues and Practice, 25* (4), 6-20.

MISLEVY, R.J., STEINBERG, L.S., AND ALMOND, R.G. 2003. On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

MISLEVY, R., STEINBERG, L., ALMOND, R., AND LUKAS, J. 2006. Concepts, Terminology, and Basic Models of ECD. *Automated Scoring of Complex Tasks in Computer-Based Testing*, D. WILLIAMSON, R. MISLEVY, AND I. BEJAR, Eds. Lawrence Erlbaum Associates, Mawah, NJ.

MITROVIC, A., MAYO, M., SURAWEERA, P., AND MARTIN, B. 2001. Constraint-Based Tutors: A Success Story. In *Proceedings of the 14th International Conference on Industrial and Engineering Application of Artificial Intelligence and Expert Systems: Engineering of Intelligent Systems, IEA/AIE-2001. LNCS 2070*, L. MONOSTORI, J. VANCZA, AND M. ALI, Eds. Springer-Verlag, Budapest, Hungary, 931-940.

MONTALVO, O., BAKER, R., SAO PEDRO, M., NAKAMA, A., AND GOBERT, J. 2010. Identifying Students' Inquiry Planning Using Machine Learning. In *Proceedings of the 3rd International Conference on Educational Data Mining*, R. BAKER, A. MERCERON, AND P. PAVLIK, Eds. Pittsburgh, PA, 141-150.

NATIONAL RESEARCH COUNCIL. 1996. National Science Education Standards. National Academy Press, Washington, D.C.

NATIONAL RESEARCH COUNCIL. 2011. A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. National Academy Press, Washington, D.C.

NEWELL, A. 1990. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

NEWELL, A., AND SIMON, H.A. 1972. *Human problem solving*. Prentice-Hall, Englewood Cliffs, NJ.

NJOO, M., AND DE JONG, T. 1993. Exploratory Learning with a Computer Simulations for Control Theory: Learning Processes and Instructional Support. *Journal of Research in Science Teaching, 30*, 821-844.

PAPERT, S. 1980. Computer-based Microworlds as Incubators for Powerful Ideas. In *The Computer in the School: Tutor, Tool, Tutee*, R. TAYLOR, Ed. Teacher's College Press, New York, NY, 203-201.

PARDOS, Z.A., HEFFERNAN, N.T., ANDERSON, B., AND HEFFERNAN, L. 2010. Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S.R. VIOLA, M. PECHENIZKIY, AND R.S.J. BAKER, Eds. Chapman & Hall/CRC Press.

PEA, R., AND KURLAND, D. 1984. On the Cognitive Effects of Learning Computer Programming. *New Ideas in Psychology, 2*, 137-168.

PELLEGRINO, J., CHUDOWSKY, N., AND GLASER, R. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academy Press, Washington, DC.

PERKINS, D. 1986. *Knowledge as design*. Erlbaum, Hillsdale, NJ.

QUELLMALZ, E., KREIKEMEIER, P., DEBARGER, A. H., AND HAERTEL, G. 2007. *A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards*. Presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April 9-13.

QUELLMALZ, E., TIMMS, M., AND SCHNEIDER, S. 2009. *Assessment of Student Learning in Science Simulations and Games*. National Research Council Report, Washington, D.C.

RAZZAQ, L., FENG, M., NUZZO-JONES, G., HEFFERNAN, N.T., KOEDINGER, K.R., JUNKER, B., RITTER, S., KNIGHT, A., ANISZCZYK, C., CHOKSEY, S., LIVAK, T., MERCADO, E., TURNER, T.E., UPALEKAR. R, WALONOSKI, J.A., MACASEK. M.A., AND RASMUSSEN, K.P. 2005. The Assistment Project: Blending Assessment and Assisting. In *Proceedings of the 12th Artificial Intelligence In Education*, C.K. LOOI, G. MCCALLA, B. BREDEWEG, AND J. BREUKER, Eds. ISO Press, Amsterdam, The Netherlands, 555-562.

REIMANN, P. 1991. Detecting functional relations in a computerized discovery environment. *Learning and Instruction, 1* (1), 45-65.

RESNICK, M. 1997. *Turtles, Termintes, and Traffic Jams: Explorations in Massively Parallel Microworlds.* MIT Press, Cambridge, MA.

REYE, J. 2004. Student modeling Based on Belief Networks. *International Journal of Artificial Intelligence in Education, 14* (1), 1-33.

RICHARDSON, J. 2008. *Science ASSISTments: Tutoring Inquiry Skills in Middle School Students.* Unpublished Interactive Qualifying Project, Worcester Polytechnic Institute, Worcester, MA.

RITTER, S., HARRIS, T., NIXON, T., DICKINSON, D., MURRAY, R. C., AND TOWLE, B. 2009. Reducing the Knowledge-Tracing Space. In *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009*, T. BARNES, M. DESMARAIS, C. ROMERO, AND S. VENTURA, Eds. Cordoba, Spain, 151-160.

ROMERO, C., AND VENTURA, S. 2010. Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 40* (6), 601-618.

ROWE, J., AND LESTER, J. 2010. Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. In *Proceedings of the 6th Annual AI and Interactive Digital Entertainment Conference, AIIDE 2010*, C.G. YOUNGBLOOD AND V. BULITKO, Eds. AAAI Press, Palo Alto, CA, 57-62.

RUIZ-PRIMO, M., AND SHAVELSON, R. 1996. Rhetoric and reality in science performance assessment. *Journal of Research in Science Teaching, 33* (10), 1045-1063.

RUPP, A.A., GUSHTA, M., MISLEVY, R.J., AND SHAFFER, D.W. 2010. Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment, 8* (4), 3-47.

RUPP, A.A., LEVY, R., DICERBO, K.E., SWEET, S., ET AL. This issue. Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*.

RUSSELL, S., AND NORVIG, P. 2009. *Artificial Intelligence: A Modern Approach, 3rd Edition.* Prentice Hall, Upper Saddle River, NJ.

SAO PEDRO, M., BAKER, R., GOBERT, J., MONTALVO, O., AND NAKAMA, A. 2011. Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction,* DOI: 10.1007/s11257-011-9101-0.

SAO PEDRO, M.A., BAKER, R.S., MONTALVO, O., NAKAMA, A., AND GOBERT, J.D. 2010. Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. In *Proceedings of the 3rd International Conference on Educational Data Mining*, R. BAKER, A. MERCERON, AND P. PAVLIK, Eds. Pittsburgh, PA, 181-190.

SCALISE, K., TIMMS, M., CLARK, L., AND MOORJANI, A. 2009. *Student learning in science simulatons: What makes a difference?* Paper presented at the America Educational Research Association, San Diego, CA.

SCHAUBLE, L., GLASER, R., DUSCHL, R.A., SCHULZE, S., AND JOHN, J. 1995. Students' Understanding of the Objectives and Procedures of Experimentation in the Science Classroom. *The Journal of the Learning Sciences, 4*, 131-166.

SCHUNN, C.D., AND ANDERSON, J.R. 1998. Scientific Discovery. In *The Atomic Components of Thought*, J.R. ANDERSON, Ed. Lawrence Erlbaum Associates Inc., Mahwah, NJ, 385-428.

SCHUNN, C.D., AND ANDERSON, J.R. 1999. The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23* (3), 337-370.

SHAVELSON, R., WILEY, E.W., AND RUIZ-PRIMO, M. 1999. Note On Sources of Sampling Variability in Science Performance Assessments. *Journal of Educational Measurement, 36* (1) , 61-71.

SHUTE, V., AND GLASER, R. 1990. A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 55-71.

STEVENS, R., SOLLER, A., COOPER, M., AND SPRANG, M. 2004. Modeling the Development of Problem Solving Skills in Chemistry with a Web-Based Tutor. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems, ITS 2004. LNCS 3220*, J.C. LESTER, R.M. VICARI, AND F. PARAGUACU, Eds. Springer, Maceio, Alagoas, Brazil, 580-591.

TSIRGI, J. 1980. Sensible Reasoning: A Hypothesis about Hypotheses. *Child Development, 51*, 1-10.

VAN JOOLINGEN, W., AND DE JONG, T. 1991. Supporting Hypothesis Generation by Learners Exploring an Interactive Computer Simulation. *Instructional Science, 20*, 389-404.

VAN JOOLINGEN, W.R., AND DE JONG, T. 1997. An extended dual search space model of scientific discovery learning. *Instructional Science, 25* (5), 307-346.

VAN JOOLINGEN, W.R., AND DE JONG, T. 2003. SimQuest, Authoring Educational Simulations. *Authoring Tools for Advanced Technology Learning Environments: Toward Cost-effective Adaptive, Interactive, and Intelligent Educational Software*, T. MURRAY, S. BLESSING, AND S. AINSWORTH, Eds. Kluwer, Dordrecht, The Netherlands, 1-31.

WALONOSKI, J., AND HEFFERNAN, N. 2006. Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006. LNCS 4053*, M. IKEDA, K. ASHLAY, AND T.-W. CHAN, Eds. Springer-Verlag, Johngli, Taiwan, 382-391.

WENGER, E. 1987. *Artificial Intelligence and Intelligent Tutoring of Knowledge*. Morgan Kaufmann, Los Altos, CA.

WHITE, B., AND FREDERIKSEN, J. 1998. Inquiry, Modeling and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction, 16* (1), 3-118.

WILLIAMSON, D., MISLEVY, R., AND BEJAR, I. 2006. *Automated Scoring of Complex Tasks in Computer-Based Testing*. Lawrence Erlbaum Associates, Mahwah: NJ.

WITTEN, I., AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, CA.