Running head: DIGITAL ASSESSMENT ENVIRONMENTS FOR SCIENTIFIC INQUIRY

Word Count: 13,428

Ciation: Gobert, J.D., & Sao Pedro, M.A. (2017). Digital assessment environments for scientific inquiry practices. In Rupp, A.A. & Leighton, J.P (Eds.) The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications. West Sussex, UK (pp. 508-534).

Digital Assessment Environments for Scientific Inquiry Practices

Janice D. Gobert
Graduate School of Education
Rutgers University and Apprendis LLC

Michael A. Sao Pedro
Apprendis LLC

Abstract

In this chapter, we provide an overview of the design, data-collection, and data-analysis efforts for a digital learning and assessment environment for scientific inquiry / science practices called *Inq-ITS* (*I*nquiry *I*ntelligent *T*utoring *S*ystem; www.inqits.org). We first present a brief literature review on current science standards, learning sciences research on students' difficulties with scientific inquiry practices, and modern assessment design frameworks. We then describe how we used pilot data from four case studies with hands-on inquiry tasks for middle school students to better understand these difficulties and design various components of the Inq-ITS system to support students' inquiry accordingly. Lastly, we describe how we used key computational techniques from knowledge-engineering and educational data mining to analyze data from students' log files in this environment to (1) automatically score students' inquiry skills, (2) provide teachers with fine-grained, rich, classroom-based formative assessment data on these practices, and (3) react in real time to scaffold students as they engage in inquiry.

Key words: Digital assessment environment, scientific inquiry practice, Inq-ITS, intelligent tutoring system, educational data mining.

Digital Assessment Environments for Scientific Inquiry Practices

Despite the billions of dollars spent on education every year in the U.S., the expenditures do not result in superior test results for American students (Hanushek, 2005). Specifically, American students continue to underperform in science compared to other developed countries. For example, in 2013, the United States ranked 21st worldwide on a key educational survey called the *Program for International Student Assessment* (PISA; Organization for Economic Co-operation and Development, 2014). There are a few reasons contributing to the poor test scores of American students on international comparisons of science competency such as PISA.

First, the current public school system was modeled on factories with a "one-size fits all" approach to teaching (Christensen, Horn, & Johnson, 2008). This approach does not recognize the various dimensions on which students differ, including but not limited to: prior content knowledge, skills to conduct science inquiry, epistemological understanding of science, and engagement and/or motivation for science learning, all of which influence school performance.

Second, standardized tests, which typically use multiple-choice and fill-in-the-blank items that tap rote science "facts" as a measure of science content knowledge are not measuring the knowledge and competencies proposed by national frameworks such as the *Next Generation Science Standards* (NGSS; NGSS Lead States 2013). Key competencies required by the NGSS include, for example, asking questions, planning and carrying out experiments, analyzing and interpreting data, warranting claims with evidence, and communicating findings (Clarke-Midura, et al., 2011; deBoer et al., 2008; Haertel, Lash, Javitz, & Quellmalz, 2006; Quellmalz & Haertel, 2004; Quellmalz, Kreikmeier, DeBarger, & Haertel, 2007).

The standardized, multiple choice tests developed years ago are no longer sufficient as a measure of 21st century skills and knowledge as described in the NGSS, which include process skills for 'doing science' (i.e., science practices) as critical aspects of science literacy (Perkins, 1986) so that people can apply and transfer their knowledge in flexible ways (NGSS Lead States, 2013). That is, these tests do not provide basic information about higher-order thinking in science such as students' processes and reasoning (Leighton & Gierl, 2011). As discussed elsewhere (Gobert et al., 2013), the limitations of these multiple choice tests are in part an artifact of a simplified conceptualization of what constituted science understanding at the time these accountability tests were designed (diCerbo & Behrens 2012; Mislevy et al., 2012).

A third but related barrier to cultivating scientifically literate students is the lack of systems that can provide individualized, real time scaffolding of students' science practices. From multiple choice tests designed for accountability purposes, educators cannot know who needs help and, as a result, any kind of feedback to students that is needed for deep learning from these tests is given too late to be formative - typically months after the school year has ended. As a result, many students struggle in silence as confirmed by prior literature and consistent data showing poor motivation for and disengagement from science learning (see, e.g., Gobert, Baker, & Wixon, 2015). Moreover, due to the fact that science inquiry is an ill-defined task, there are a myriad of ways in which students conduct inquiry, both when they are on the "right" track and when they are not (Kuhn, 2005). Although progress on real time systems for well-defined domains like math and computer science has been made (cf., Koedinger & Corbett, 2006; Corbett & Anderson, 1995), there are few to no systems that adapt to individual learners as they conduct science inquiry.

These assessment challenges and needs for adaptive instruction have led to the development of new technology-centered measurement paradigms for science education (see Timms, Clements, Gobert, Ketelhut, Lester, Reese, & Wiebe, 2012). At present, key organizations such as the *National Center for Education Statistics* that is responsible for the *National Assessment of Educational Progress* (NCES, 2011), the *Organisation for Economic Co-operation and Development* that is responsible for PISA (OECD, 2014), the *National Educational Technology Plan*, and the *National Research Council* (NRC, 2011) all acknowledge the benefits and potential of technology-based systems for the assessment of science inquiry (see chapter by Oranje, this volume). Computer-based environments like the one that we describe in this chapter along with others (see, e.g., Quellmalz et al, 2012; Ketelhut & Dede, 2006; Clarke et al, 2012; Leeman-Munk, Wiebe, & Lester, 2013; Timms, et al., 2012) are providing new possibilities for assessing science and are now being considered as alternatives to traditional assessments of inquiry

(Behrens, 2009; Gobert et al., 2012, 2013; Pellegrino, Chudowski, & Glaser, 2001; Quellmalz et al., 2009).

In the following, we provide an overview of the design, data-collection, and data-analysis efforts for our digital assessment environment for scientific inquiry / science practices called *Inq-ITS* (*I*nquiry *I*ntelligent *T*utoring *S*ystem; www.inqits.org).
Inq-ITS is an example of a system that was designed and instrumented to generate assessments of students' science practices as they engage in rich, authentic inquiry tasks. Inq-ITS was explicitly designed to prioritize the assessment of inquiry rather than the learning of science, which many other curriculum-focused inquiry systems emphasize (cf., Linn & Hsi, 2000). Our work builds on the inquiry and assessment research of others and forges new ground for inquiry assessment and scaffolding of science with its application of data mining techniques (Gobert et al, 2013).

We have organized this chapter into three main sections as follows. In the first section we review key literatures on science learning policy, students' difficulties with science inquiry, and assessment design to provide a background for the design of the Inq-ITS system. In the second section, we describe how we used pilot data from four case studies with hands-on inquiry tasks for middle school students to better understand these difficulties and design various components of the Inq-ITS system to support students' inquiry accordingly. In the third section, we describe how we used key computational techniques from knowledge-engineering and educational data mining to analyze data from students' log files to (1) automatically score students' inquiry skills, (2) provide teachers with fine-grained, rich, classroom-based formative assessment data on these practices, and (3) react in real time to scaffold students as they engage in inquiry.

Foundations for the Design of the Inq-ITS System

NGSS

The NGSS, as the National Research Council's new framework for K-12 Science Education, emphasize content learning as well as inquiry practices, as did its predecessors (see NSES, 1996). However, in the newest framework greater emphasis is placed on the rich integration of authentic practices in science with disciplinary content knowledge so that students will possess well-honed learning strategies that can be transferred in more flexible ways. These inquiry practices are:

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence, and
8. Obtaining, evaluating, and communicating information

The NGSS also prescribe that American middle school students develop content understanding of topics as shown in Table 1.

[INSERT TABLE 1 ABOUT HERE]

Finally, the NGSS also describe six cross-cutting science concepts:

1. Cause and effect
2. Scale, proportion, and quantity
3. Systems and system models
4. Energy and matter
5. Structure and function, and
6. Stability and change.

Given the importance placed on rich science inquiry practices that are aligned to the needs of the 21st century and the poor performance in science by American students in studies such as PISA, there follows a need for better assessments that can provide more fine-grained data about ‘how’ students are learning (or ‘not’ learning, as appears to be the case for many students).

Learning Sciences

Models for learning. There were several related bodies of literature that we drew on in the design of our system specifically. Briefly, the main literature includes: *causal models* (White & Frederiksen, 1990; Schauble et al, 1991; Raghavan & Glaser, 1995), *visualization generation and comprehension* (Gobert, 1994; Gobert & Frederiksen, 1988; Gobert, 2005; Kindfield, 1993; Larkin & Simon, 1987; Lowe, 1989), *mental models* (Gentner & Stevens, 1983; Johnson-Laird, 1983) and *model-based learning* (Gobert & Buckley, 2000; Harrison & Treagust, 2000), as well as the vast body of literature on students' alternative conceptions (Pfundt & Duit, 1988; Driver, 1983) and difficulties with inquiry (cf., Kuhn, 2005).

Most relevant to the theoretical framework that undergirds our system is how people learn with rich visual representations, specifically model-based learning; see Figure 1 for a graphical representation of this framework. The design of our system, its scaffolds, and other assessment components, are based on model-based teaching and learning (Gilbert, J. 1993; Gilbert, S., 1991; Gobert & Buckley, 2000). This, in turn, prescribed the design of the interface as well as the design of the tools and widgets in Inq-ITS, which we describe further below.

[INSERT FIGURE 1 ABOUT HERE]

Briefly, model-based learning (Clement, Brown, & Zietsman, 1989; Gobert & Buckley, 2000; Gobert & Clement, 1999; Harrison & Treagust, 2000) is a theory of science learning that integrates basic research in cognitive psychology and science education. The tenets of model-based learning are based on the presupposition that deep understanding requires the construction of mental models of the phenomena under study, and that all subsequent problem-solving, inference making, or reasoning are done by "running" and manipulating these mental models (Johnson-Laird, 1983). We view mental models as internal cognitive representations used in reasoning (Brewer, 1987; Rouse & Morris, 1986); thus, we define model-based learning as a dynamic, recursive process of learning by constructing mental models of the phenomenon under study. It involves the formation, testing, and subsequent reinforcement, revision, or rejection of those mental models (Gobert & Buckley, 2000; Clement, 1993; Stewart & Hafner, 1991). This is analogous to hypothesis development and testing seen among scientists (Clement, 1989) and also, we argue, a form of reasoning used in conducting inquiry.

In our theoretical framework, we also use D. Norman's contemporary definition of an *affordance* (Norman, 1983). That is, the ways in which one learns from a visual representation (i.e., constructs a mental model) and the features that the *microworld* affords the learner are dependent on the learner's knowledge, skills, predispositions, and other characteristics. This is a more contemporary interpretation of the original use of the term affordance by the perceptual psychologist J.J. Gibson, who claimed that an affordance of a representation (or object) is independent of the user's knowledge, skills, predispositions, and other characteristics (Gibson, 1977). Thus, the prior knowledge, epistemological frameworks, inquiry skills, and other variables related to the learner that s/he brings to bear when engaging in inquiry with the representation will play a role in the nature of the resulting mental model (Gentner & Stevens, 1983; Johnson-Laird, 1983).

The notion of what features provide affordances and for whom affected both our interface and scaffolding design. That is, since many students lack the necessary domain knowledge to guide their search processes through diagrams/models during learning (Lowe, 1989; Gobert, 1994; Gobert & Clement, 1999), our system and its scaffolds need to support learners so that their inquiry can be productive, resulting in the construction of rich mental models with which they can engage in sophisticated model-based reasoning.

Students' difficulties with scientific inquiry. As part of our design, we needed to know what difficulties middle school students have when conducting inquiry. We conducted a thorough literature review on students' difficulties with inquiry to better understand the nature of these. This helped us to both concretize the sub-skills underlying the inquiry practices identified in the NGSS, as well as design tools and widgets to guide students in conducting inquiry.

Many studies have shown that students have difficulty with inquiry learning in general. Students do not plan which experiments to run (Glaser et al., 1992) and can act randomly (Schauble, Glaser, et al., 1991). They have difficulty setting goals (Charney, Reder, Kusbit, 1990), monitoring their progress (de Jong et al., 2005; de Jong, 2006), and recording their progress (Harrison & Schunn, 2004). Furthermore, they concentrate more on executing procedures than what they might learn or infer from experimenting, and if the data-gathering process is lengthy, they lose track of why they are collecting data

(Krajcik et al., 2000). Students have also been shown to demonstrate some difficulties in terms of certain very specific inquiry skills targeted in the national frameworks for which we sought to design assessment metrics and scaffolds. We discuss findings from prior research on five skills that were critical to the design of Int-ITS briefly in the following.

First, when 'generating hypotheses' - referred to as asking questions in the NGSS - students may have difficulties choosing which variables to work with (Chinn & Brewer, 1993; Klahr & Dunbar, 1988; Kuhn et al., 1995), including identifying the proper independent variable (Richardson, 2008). They may also have difficulty translating and understanding how theoretical variables and manipulable variables relate to each other (van Joolingen & de Jong, 1997; Glaser et al., 1992).

Second, when 'planning and carrying experiments', students may not test their articulated hypotheses (van Joolingen & de Jong, 1991, 1993; Kuhn, Schauble, Garcia-Mila, 1992; Schauble, Klopfer, Raghavan, 1991) or may gather insufficient evidence to test hypotheses (Shute & Glaser, 1990; Schauble, Glaser et al., 1991) by running only one trial (Kuhn, Schauble, Garcia-Mila, 1992) or running the same trial repeatedly (Kuhn, Schauble & Garcia-Mila, 1992; Buckley, Gobert & Horwitz, 2006). They may also change too many variables (Glaser et al., 1992; Reimann, 1991; Tschirgi, 1980; Shute & Glaser, 1990; Kuhn, 2005; Schunn & Anderson, 1998, 1999; Harrison & Schunn, 2004; McElhaney & Linn, 2008, 2010), may run experiments that try to achieve an outcome (e.g., make something burn as quickly as possible), or may design experiments that are enjoyable to execute or watch (White, 1993), as opposed to actually testing a hypothesis (Schauble, Klopfer & Raghavan, 1991; Schauble, Glaser, Duschl, Schulze & John, 1995; Njoo & de Jong, 1993).

Third, when 'analyzing and interpreting data' students may show confirmation bias (i.e., they will not discard a hypothesis based on negative results) (Klayman & Ha, 1987; Dunbar, 1993; Quinn & Alessi, 1994; Klahr & Dunbar, 1988; Dunbar, 1993). They may draw conclusions based on confounded data (Klahr & Dunbar, 1988; Kuhn, Schauble & Garcia-Mila, 1992; Schauble, Glaser, Duschl, Schulze & John, 1995), they may not relate outcomes of experiments to theories being tested (Schunn & Anderson, 1999), and they may reject theories without disconfirming evidence (Klahr & Dunbar, 1988). They may have difficulty linking data back to hypotheses (Chinn & Brewer, 1993;

Klahr & Dunbar, 1988; Kuhn et al, 1995), may have difficulty interpreting data displays like graphs, or may have difficulty interpreting important differences between related variables (e.g., time to evaporate vs. rate of evaporation) shown in data displays (cf. McDermott, Rosenquist, & van Zee, 1987).

Fourth, when ‘constructing explanations’*,* students may be overly reliant on theoretical arguments as opposed to evidence (Kuhn, 1989, 1991; Kuhn, Katz & Dean, 2004; Ahn et al., 1995; Ahn & Bailenson, 1996; Brem & Rips, 2000; Schunn & Anderson, 1999), they may struggle to provide appropriate evidence for their claims (McNeill & Krajcik, 2007), or they may analyze data so as to protect prior beliefs, which can lead to faulty causal attribution (Kuhn et al., 1995; Keselman, 2003; Kuhn & Dean, 2004).

Finally, when ‘communicating findings’*,* students may have difficulty articulating and defending claims (Sadler, 2004). They may tend to focus on what they did as opposed to what they found out, may not link data and conclusions, and may not relate results to their own knowledge/questions (Krajcik, et al., 1998). They may also struggle to provide reasoning to describe why evidence supports claims (McNeill & Krajcik, 2007).

Assessment System Design

Intuitively speaking, it makes sense to design educational assessments based on the vast literature from the learning sciences summarized in “How People Learn” (Bransford, Brown, & Cocking, 2000). Briefly, the literature in that volume spans from the onset of the information-processing perspective on learning circa 1960 to present day and has been extremely informative regarding the role of prior knowledge in learning, the nature of mental models and their role in reasoning, as well as domain-specific content learning and teaching, including science (Duschl, Schweingruber, & Shouse, 2007).

In a white paper report that extended “How People Learn”, Pellegrino (2009) outlined some key principles that are very noteworthy in guiding the design of assessments of NGSS practices so that the resulting data can be used to “educate and improve student performance, rather than merely to audit it” (Wiggins, 1998, p. 7). In brief, these principles emphasize that assessments should be integrated with curricular/instructional needs including domain-subject matter learning (for science, this includes content and practices) and that assessments need to be framed by current theories and data about student cognition and learning, including learning progressions and expert-novice differences, in addition to students' prior knowledge.

From a system designer's perspective, the practical assessment problem becomes the following: how do we take the policy documents about what science literacy is and what we expect students to be able to know and do in science and use these to inform the design and development of a valid and reliable system capable of generating performance-based assessments? How do we design a system that is scalable to large numbers of users so that science literacy on a broad scale can be realized? It is clearly necessary that the new assessment permit better inferences about students' knowledge, skills, and inquiry practices when compared to more traditional multiple choice tests , while still providing evidence of both validity and reliability, while being neither too expensive nor too laborious to construct.

Properties of assessment systems. Leighton & Gierl (2011) offer three indices for evaluating assessment items/systems, namely *granularity*, *measurability*, and *instructional relevance*. Briefly, granularity refers to the depth and breadth of the knowledge and skills being measured by the system. Specifically, to permit inferences about what students know, underlying cognitive models must be described/collected/reported at a level of specificity that will provide meaningful information about students' performance so that teachers (or the system itself) can provide necessary feedback. If developed at the right level of granularity, a teacher can, in turn, use these formative data to inform their instruction. Alternatively, support in the form of scaffolding can be done in real time via an automated pedagogical agent as in the Inq-ITS system that we describe in this chapter.

The second criterion for evaluating an assessment is its measurability in order to link learning with assessment. Specifically, the knowledge and skills in the cognitive model must be described in a way that would allow a developer to create a test item or task to measure that particular knowledge or skill. Later in this chapter we describe how we used information from four case studies to develop articulations of the knowledge and skills for scientific inquiry practice for our Inq-ITS system.

The third criterion for evaluating assessments is instructional relevance. That is, in developing a cognitive model, the knowledge and skills must be instructionally relevant and meaningful to the relevant group of educational stakeholders such as teachers, superintendents, and policy-makers. For example, teachers need highly actionable data that are easy to understand and use in real time to inform instruction (Huff & Goodman, 2007). Instructional relevance is generally related to grain size. For example, when data are derived from students' logs - as is the case in with our Inq-ITS system - the data must be aggregated from their finer-grained level up to a level that is instructionally relevant for teachers for the purposes of instruction and scaffolding.

Evidence-centered design. Despite the rich theoretical frameworks from the learning sciences and vast amount of findings on how people learn science, the development of resources to assess science is lagging behind (Leighton & Gierl, 2011). Specifically, the broad and deep literature base from the learning sciences about what students know and the types of knowledge they use in reasoning should be used to guide the design of test items and tasks for assessment. If designed in this way, there is greater potential to strengthen validity arguments regarding the inferences that can be made about students' knowledge from such items (Leighton & Gierl, 2011).

Delving a level deeper in terms of its specificity for guiding the development of assessments for science in particular, Mislevy and colleagues (e.g., Mislevy et al., 2012) thus proposed the *evidence-centered design* (ECD) framework. This framework describes how the analysis of key practices in a domain can be used to inform the design of assessments for that domain. Domain analysis is similar in spirit to task analysis as described in the information-processing literature (Newell & Simon, 1972), but ECD is has the explicit goal of assessment design, whereas task analysis is typically used to characterize learning (Newell, 1990).

In calibrating a system for assessment purposes, Mislevy et al (2012) explicitly state how domain analysis and subsequent domain modeling processes are used to inform the design of the *conceptual assessment framework* within the ECD framework. In the context of Inq-ITS, scientific inquiry practices to be assessed are specified in a *student model* and then connected to a *task model* that specifies features of tasks as well as questions that would elicit the evidence of learning. Observable behaviors then result in indicators that are used for evidence identification and accumulation processes within an *evidence model* that specify the nature of student responses that indicate levels of proficiency; for a full description of how our ECD models were derived see Gobert et al. (2012).

We derived our cognitive model for the sub-skills underlying the inquiry practices from our think-aloud data from the four case studies, which we discuss in the next section, and the previously reviewed literature on students' difficulties with scientific inquiry. In Inq-ITS, the task model includes the activities conducted in the microworld that reveal students' proficiencies for each sub-skill of interest. Finally, the evidence model specify how one uses work products (i.e., end-state products) and processes (i.e., actions/behaviors as indicated in their log files) to assess students' inquiry practices. These data are then aggregated and analyzed to yield performance indicators that are used as evidence of students' proficiencies for each inquiry practice and their respective sub-skills.

To create automated evidence-based assessment summaries within a complex system like Inq-ITS, it is critical to know how to leverage modern computational techniques in order to analyze the rich log file data that are generated and captured as students engage in scientific inquiry tasks. Although there are many on-line learning environments for science, few are leveraged to assess the skills that they were designed to foster (Quellmalz et al., 2009). Computational techniques adapted or adopted from domains such as *computer science* or *educational data mining* in particular are necessary to handle the analysis of data both in terms of the grain-size and volume of log files that are generated in rich interactive systems (Behrens, 2013; Gobert et al, 2013; Mislevy et al, 2012).

In line with ECD thinking, it is critical to concretize the sub-skills underlying the inquiry practices at a level of granularity that informs decisions about what kinds of data along with what kinds of computational techniques to generate assessment metrics of inquiry practices are needed. This design work needs to be done before the computational techniques can be developed for assessment rather than post hoc once the environments are already created. In the third section we describe how we analyzed our data by leveraging both knowledge-engineering and educational data mining techniques. The resulting computational techniques were designed to handle both the fine grain size of our data and its large volume in order to assess students and scaffold them in real time (Gobert et al., 2012, 2013 provides a thorough description of this approach).

Case Studies with Think-aloud Components

Using the above literatures as a basis, we designed and conducted a series of one-on-one case studies with students using *think aloud protocols* (Richardson, 2008). Think-aloud protocols are assumed to present a "trace" of the learner's cognitive processes in that the object being described as a person thinks out loud is assumed to be information/ knowledge that is currently being attended to in execution of the task (Ericsson & Simon, 1980). Methods used to analyze think aloud data can be very fine-grained. Some methods of protocol analysis are done at the propositional (i.e., basic idea unit) level (see Frederiksen, 1975; 1986) or at the clause level (Chi, 1997), providing key information about the semantic units underlying thinking. Think-aloud data have been used to provide information about particular facets of task performance that can be used to develop canonical models for software development (Ericsson & Simon, 1980).

In order to inform the initial design of our Inq-ITS environment, four case studies with think-aloud components were conducted in our partner middle schools. Across the four case studies, we sought to (1) characterize how middle school students naturally approach scientific inquiry tasks, (2) develop a set of scaffolds for inquiry to be integrated into a technical environment, and (3) determine the effectiveness of various prompts and scaffolding tasks at fostering inquiry practices.

Case Study 1

The first case study was designed to characterize how students naturally approach an inquiry task, to get a sense of what the students already understood in terms of inquiry, and to better understand common areas of weakness with respect to both inquiry in general as well as areas of weakness within designing controlled experiments specifically (Chen & Klahr, 1999). In this case study, fourteen randomly selected middle school students were selected from a range of class levels including high, average, and lower-performing students. Each student was tested on an individual basis. Students were presented with a physical ramp apparatus whose features (steepness, run length, the type of ball, and surface) could be changed. They also were given blank pieces of paper and a pencil to record their findings.

Students were first asked which features they thought they could change on the ramp that would affect how far the ball would roll. They were then told that these features were called variables. Next, they were asked to state a hypothesis and run an experiment to test their hypothesis on how the steepness of the ramp affects how far the ball would roll. After running the experiment, the students were asked to explain, based on their data, how steepness affects how far the ball rolls. When the students finished testing how steepness affects how far the ball rolled, they moved on to test the effect of run length, the type of ball, and surface again on how far the ball rolled. Then they were asked to reflect back on the first experiment and say what they would have done differently if they were to run it again. This question was asked to test if they had acquired meta-knowledge about how to conduct controlled experiments.

The prompts in this case study, which were intended to be a “gentle guide” toward improving the student’s strategies, were not prewritten. If a student continued to use the same type of inquiry strategy, the next prompt given was slightly more direct. For example, if a student was demonstrating “buggy” inquiry, the student was given a prompt and then given more time to interact with the apparatus to revise his/her strategy and re-run the experiment. The approach of providing progressively more direct scaffolds was later incorporated into the Inq-ITS system.

The data from this study included all of the notes and tables made by each student during the experiment. Additionally voice data was analyzed to determine which inquiry skills the students struggled with and which prompts were helpful in improving inquiry skills. The data showed that, although students were generally fairly good at articulating a hypothesis (e.g., they included an independent variable, a dependent variable, and specified a relationship between them), they showed difficulties with other inquiry practices, many of which were previously described in the inquiry literature reviewed above.

First, students did not naturally seek to record their data and needed both prompting and a great deal of help in recording data. Specifically, many did not know how to record the data in columns with the values for each independent variable in one column and the resulting dependent variable in another column. Second, with respect to designing and conducting experiments, most students did not target one variable by changing only that one and keeping all others the same. Many students also collected data from a single trial, collected data for the same trial repeatedly, or collected data merely to reaffirm their initial hypothesis without considering alternative explanations. When describing their findings, students did not include data in their explanations; that is, they did not provide evidence for their claims with data from their table(s). It was notable, however, that with some experimenter scaffolding, students' performance was better across trials at making a table. Since recording data is a graphical literacy skill as opposed to a science inquiry skill, data on students' difficulties on recording their data led us to better understand the importance of providing an auto-populated data table for students in the design of Inq-ITS.

Case Study 2

In a second case study, we collected data from demographically-similar students attending one of our partner schools, a lower SES school in Central Massachusetts. All materials, including the physical ramp apparatus, data collection, and recording procedures were identical to case study 1. However, in this study we administered a short pretest and posttest of inquiry skills, we provided students with a lab book, and used more formalized inquiry prompts as part of the data collection procedure. Similar to study 1, we found that the students did not record findings without prompting to do so. When they had trouble with conducting controlled trials, the experimenter showed them some data in the lab book and students were able to pinpoint issues in the collection procedure of these data. However, when conducting their own trials, they typically failed to conduct controlled trials when collecting data. When analyzing data, students again struggled with writing explanations for their data. Data from this study further demonstrated the complexity and "thorniness" of students' difficulties with conducting controlled trials. Specifically, students do not tend to conduct contrasting trials sequentially, making the assessment of their knowledge of how to design controlled experiments very difficult. Later in the chapter we describe, in brief, how our algorithms do this, as well as describe the need to refine our scaffolds for this critical inquiry practice within the Inq-ITS system.

Case Study 3

In the third case study, again with the same student demographic, we used a simulated ramp environment (i.e., a virtual replica of the one students used in case studies 1 and 2). Similar to case study 1, the goal here was to see what the students did naturally in terms of inquiry but within a virtual, simulated environment. In addition, the tasks were slightly more structured than in case study 2 because we determined that a structured approach was more effective in teaching how to design controlled experiments (Klahr & Nigam, 2004). In addition, data tables that included blank rows and columns were provided to the students. Voice data and videos of students' interactions with the simulated ramp apparatus were collected and analyzed for each student.

The data from this study were very informative with respect to the breadth and depth of students' difficulties with inquiry. Specifically, students again demonstrated difficulties with recording data and all students needed to have the columns of the table set up for them by the experimenter so that they could correctly record their data. When collecting data, they repeated trials, did not collect contrasting trials, did not collect trials sequentially, and did not run controlled trials. When interpreting results, they did not attend to the appropriate data. Of interest to data interpretation, when students were asked to compare their original tables to the ones set up for them by the experimenter, the students realized that had changed too many variables, making it harder to see the effect on the variable they were trying to test. In the new tables, which were set up for them, the outcome was more salient to the students. Lastly, students had difficulties in communicating their findings in that their conclusions were not based on their data.

Case Study 4

In the fourth case study, students were drawn again from the same demographic sample. This case study was similar to Case study 3 in that the simulated ramp environment was used; however, we also used a lab book similar to Case study 2. Noting earlier results about the consistent and pronounced difficulties conducting controlled experiments, the laboratory notebook included a direct explanation about how to collect unconfounded data since it was shown that direct instruction on this skill is effective (Klahr & Nigam, 2004). This was written so that the student did not need any assistance from a human tutor; this, we deemed, would help us in beginning to formalize the prompts that would be incorporated into Inq-ITS.

Again, we found that students demonstrated problems with inquiry. Specifically, when conducting experiments, they did not collect enough data and only showed a small improvement in conducting controlled experiments. Moreover, they had problems with interpreting data and communicating findings as they had in the previous case studies. For example, even though students did not have the data needed to support their conclusion, they insisted that they had "discovered the answer" and that their data "proved it". For some students it seemed too obvious that if the ramp is steeper that the ball will roll further (e.g., one student commented that "the higher the ball, the faster it goes.") When asked if they saw those results in their table, all students answered "yes"; however, they did not give a deeper explanation of how they had demonstrated that the higher the steepness of the ramp, the further the ball will go.

Summary of Case Studies

All told, the information gathered through think-aloud activities in the case studies helped us to greatly understand the breadth, depth, and pervasiveness of students' difficulties across all of the inquiry practices outlined in the NGSS consistent with previous findings in the literature. We also used our think-aloud data to identify the level of granularity needed to conceptualize and operationalize the sub-skills underlying inquiry practices. Moreover, we used analyses of students' think-aloud data to characterize students' "natural" inquiry processes with no/minimal support (e.g., graphs, tables, widgets, scaffolds) in order to better understand students' needs for these tasks. These kinds of information identified in the hand-coding of our think-aloud protocols were valuable to the development of the algorithms needed to measure the sub-skills of inquiry and to provide appropriate scaffolds for learning within Inq-ITS. We now describe the characteristics of this system in more detail.

The Inq-ITS System

As discussed previously, Inq-ITS is a rigorous, technology-based learning environment that assesses and scaffolds middle school students as they engage in inquiry in Earth, Life, and Physical Sciences. The system can be run either in "pure assessment mode" or in "scaffolding mode", in which our virtual agent, *Rex*, jumps in to support students in real time when needed. In the following section, we describe the key features of the Inq-ITS system.

Microworlds

Inq-ITS uses *microworlds* (Papert, 1980) to engage students in scientific inquiry. Microworlds are computerized representations of real-world phenomena whose properties can be inspected and changed (Pea & Kurland, 1984; Resnick, 1997). Microworlds provide authentic inquiry opportunities because they share many features with real apparati for "doing science" (Gobert, in press), thereby providing perceptual affordances for the learner. In turn, these perceptual affordances can provide leverage for building rich conceptual knowledge (Gobert, 2005). With a microworld a learner can pose questions, plan and carry out a virtual experiment with a simulation by collecting data, then analyze their data, and then communicate their findings in the form of a scientifically warranted explanation. Inq-ITS microworlds are used for performance assessment of students' inquiry practices in that they: (a) are instrumented to log all students' interactions, (b) leverage real time analyses of log files based on knowledge-engineering and educational data mining, (c) provide assessment metrics to researchers and teachers on each inquiry skill of interest and, (d) can scaffold students' inquiry processes in real time (Gobert et al., 2013).

In developing microworlds for Inq-ITS, we surveyed the science education and learning sciences literature for students' content misconceptions for each topic across Earth, Life, and Physical Sciences to determine which variables, domain-specific properties, and domain-specific representations to include in each microworld so that students could fully engage with the content in authentic ways to more deeply understand the topic and hone their inquiry practices in these domain-specific contexts. For example, in the domain of 'state change', a common misconception in middle school is that as the amount of a substance increases, the temperature at which it will boil also increases. Mislevy et al. (2012) refer to this process as the *domain analysis* in the ECD lifecycle of assessment design and delivery; the information-processing literature refers to this more generally as *task analysis* (Newell, 1990).

An integral part of the Inq-ITS system and associated microworlds is the inclusion of inquiry widgets, as mentioned before. These widgets are important in that they scaffold students in conducting various steps of inquiry, but are also the basis upon which we collect our performance data on students' inquiry practices. Our inquiry widgets were designed in accordance with the learning sciences and science education literature on students' difficulties in conducting inquiry.

For example, the 'question asking/hypothesis' widget was designed to externalize the structure of students' hypotheses using independent and dependent variables and the relationships between them. Using this widget, students' questions/hypotheses are generated in the form of sentences. The resulting data are logged and are then used to generate metrics about the sub-skills of inquiry practices such as whether the student has included an independent variable, a dependent variable, and a relationship between them.

The 'data interpretation/analysis' widget provides a structure for the student to interpret their data after the experimental trials are completed. Similar to the hypothesis widget, the data interpretation/analysis widget presents a way for the student to create statements about the relationship between the independent and dependent variables from their trials and to warrant their claims by selecting their trials to either support or refute their hypothesis. A full description of the widgets can be found in Sao Pedro et al. (2011), and Gobert et al. (2012, 2013).

Types of Scaffolding

As described in our theoretical framework, we believe that middle school students, many of whom lack adequate prior content knowledge and have difficulties with inquiry as previously described, need guidance in conducting scientific inquiry. The degree of structure used to guide students' general and science-specific inquiry activities in learning environments is a topic that was hotly debated in the field of science education in the recent past (e.g., Kirschner, Sweller, & Clark, 2006; Hmelo-Silver, Chinn, & Duncan, 2006).

As previously mentioned, Papert's conception of inquiry with microworlds is more open-ended in terms of degree of pedagogical guidance (Papert, 1980; 1993) than inquiry with microworlds in our system. Inq-ITS allows a moderate degree of student choice, less choice than in purely exploratory learning environments (Amershi & Conati, 2009; Papert, 1980; 1993) but more choice than in classic model-tracing tutors (Koedinger & Corbett, 2006) or constraint-based tutors (Mitrovic et al., 2001). Specifically, in Inq-ITS, students' scientific inquiry is guided in three ways through (1) general scaffolding afforded by the system user interface, (2) teacher scaffolding, and (3) adaptive scaffolding by our pedagogical agent *Rex*, a cartoon dinosaur. We discuss each form of scaffolding briefly in the followng.

Inq-ITS user interface scaffolding. Students' inquiry in Inq-ITS is guided by the order in which information is provided to learners as well as by the widgets and support tools provided to learners. For example, we begin by suggesting that they develop and ask a question using a set of independent and dependent variables. Students then plan and carry out an experiment by collecting data, interpret their data, provide evidence in the form of warrants for their claims, and communicate their findings. Inq-ITS has a progress bar on the top of the screen to support them in knowing what phase of inquiry they are presently in, which is critical to students' monitoring (de Jong et al., 2005; de Jong, 2006). Additionally, the artifacts that students generate by using widgets make visible and salient both the products and processes of inquiry for the learner in order support students' meta-level understanding of inquiry.

Teacher-led scaffolding. When in pure assessment mode, assessment of students' inquiry practices is done in a stealth manner (Shute, 2011); that is, unobtrusively without taking time from instruction. Formative data collected on students' inquiry practices are provided in real time directly to teachers via an integrated assessment report shown in Figure 2, which displays information at both the class-level and individual student level. The report is generated via our knowledge-engineered rules and data mined algorithms that we describe further below.

[INSERT FIGURE 2 ABOUT HERE]

Recently, we completed the development of an alerting platform called *Inq-Blotter* (Sao Pedro, Gobert, & Betts, 2015) that automatically alerts teachers on their mobile devices as to which students are having difficulties with inquiry and on which inquiry practices and sub-skills of these. With these data literally 'in hand', teachers can walk around the room and provide assistance to students as they need it, when scaffolding is most critical to learning (Koedinger & Corbett, 2006). From these data, a teacher might decide to stop the entire class, say, if many do not understand what an independent variable is or are not conducting controlled trials, or s/he might decide to go over to help individual students in real time if there are only a few students having difficulty with a particular skill. Our reports and alerts are designed to be highly readable and to identify the students who need most help for the teacher while they monitor the overall class' performance. These reports and alerts were designed in collaboration between our user experience designer, graphic designers, and our partner teachers using an iterative design process of conceptual sketches and mock-ups. We then interviewed teachers as to the usability of these reports and the levels of aggregation that they needed in the reports and tweaked the reports as needed.

Automated adaptive scaffolding via Rex. Assessment can be done either without *Rex* or with *Rex*. In the full embodiment of our system with *Rex*'s scaffolding capacity, our goal was to provide the optimal degree of guidance so that students' inquiry skills could be honed in real time, when real time feedback is most effective (Koedinger & Corbett, 2006). By running Inq-ITS in scaffolding mode, assessment is seamlessly integrated with instruction so that skills can be developed and assessed in the rich contexts in which they are developing (Mislevy et al., 2003).

Rex, our pedagogical agent, provides scaffolds on a particular inquiry practice when - and only when - our system detects that the student needs this type of help via our data-mined algorithms (for a fuller description see Gobert et al., 2013). In this way, Vygotsky's notion of scaffolding within the *zone of proximal development* (1978) can be realized. This automated scaffolding approach is used instead of on-demand help in which students explicitly ask for support (e.g., Anderson et al., 1995) because on-demand help requires metacognitive knowledge (Aleven & Koedinger, 2000; Aleven, McLaren, Roll & Koedinger, 2004). Given that students have difficulty monitoring their progress (de Jong et al., 2005), we had empirical evidence to believe that students would be unaware when they are in need of help during inquiry.

Inq-ITS has four types of automated scaffolds embodied by *Rex*: (1) *orienting scaffolds* to help students monitor where they are in the inquiry process (de Jong et al., 2005), (2) *conceptual scaffolds* to provide students conceptual information needed for the current task (e.g., *Rex* may explain why controlled experiments are important for testing a hypothesis), (3) *procedural scaffolds* to help students on the current task (e.g., *Rex* may instruct the student to "construct a controlled trial, relative to your last trial run" or to "design experiments by changing only one variable while keeping the others the same"), and (4) *instrumental scaffolds* to provide the student direct instruction as to what to do on the current task (e.g., *Rex* may break down strategic help in a step by step fashion, providing a type of worked example from which to learn) (Koedinger & Aleven, 2007).

In short, to us, the goal of providing scaffolded support for inquiry practices via the teacher or *Rex* in the form of orienting messages or conceptual/procedural strategies is not equivalent to direct instruction of formulas and rote science facts as characterized in Kirschner et al (2006). In Inq-ITS, we scaffold students' inquiry practices since (1) these are not likely to develop naturally (Kuhn, 2005); (2) students can become lost and frustrated and their confusion can lead to misconceptions if they are not scaffolded (Brown & Campione, 1994), (3) teachers spend considerable time scaffolding students' procedural skills (Aulls, 2002), and (4) there are many "lost" opportunities for learning and assessments if students are not guided properly via scaffolds (e.g., if students are not testing their hypothesis or if their data are confounded, all subsequent inquiry tasks such as data interpretation, warranting claims, and developing explanations are moot since their data do not afford the possibility of successfully completing these tasks due to "buggy" data collected during data collection phase of inquiry).

Illustrative Vignette

To make the previous ideas more concrete, we now present a small vignette in the context of a 'states of matter' microworld that students use to conduct inquiry to determine the effects of the independent variables (e.g., level of heat, amount of substance) on the dependent variables (e.g., time to melt, temperature when melted); see Figure 3 for a screenshot of this microworld.

[INSERT FIGURE 3 ABOUT HERE]

After the student has had some time to explore the microworld, the student uses the 'question asking/hypothesis' widget to generate a hypothesis. When the student finishes this, it is checked for correctness using a knowledge-engineered rule (Feigenbaum & McCorduck, 1983). If the student incorrectly enters a dependent variable in place of an independent variable, this particular sub-skill (i.e. to distinguish independent from dependent variables) is auto-scored as 'incorrect' and the teacher report shown earlier in Figure 2 will be auto-populated with this information so the teacher will always know the status of his/her students' inquiry skills; such timely feedback is critical to deep learning.

Students then design and conduct their experiment by first selecting variables to manipulate and then running a simulation. This provides students additional opportunities to demonstrate their understanding of independent and dependent variables as well as their understanding of how to test a hypothesis using controlled trials. As students collect data by running trials with the simulation, it is highly likely that some students will not design controlled experiments whereby one variable is systematically targeted across trials and all other variables are held constant (Chen & Klahr, 1999). Once a number of trials have been completed, our data-mined assessment algorithm is able to assess whether students are successfully demonstrating this skill (Sao Pedro, Baker, & Gobert, 2012). If not, this information is automatically updated in the teacher report. Again, as the teacher helps one student, our algorithms continually update the report/alerts.Once sufficient data are gathered to support or refute the hypothesis, students interpret their data and warrant their claims using the 'analysis interpretation' widget. Again, as in hypothesis formation, a knowledge-engineered rule checks the student's interpretation both for correctness, and whether they have selected the correct data to warrant their claim. The report indicates which students are most in need of help on which skill(s).

Generation of Assessment Metrics

The log files of students' actions collected unobtrusively and in situ within the Inq-ITS system provide a fertile basis upon which to generate performance-based assessments of rich inquiry processes (Clarke-Midura, Dede, & Norton, 2011). Additionally, the resulting evidence about key student competencies from log files can be connected to the evidence from the artifacts or products they create as a result of the captured activities (see, e.g., Rupp et al., 2010).

As alluded to previously, our system assesses inquiry practices using a combination of knowledge-engineered rules (Feigenbaum & McCorduck, 1983) and data mined algorithms (Romero & Ventura, 2007; Baker & Yacef, 2009), depending on whether the inquiry practice of interest is more well-defined or more ill-defined. Knowledge-engineering techniques (see Shute & Glaser, 1990; Schunn & Anderson, 1998), which work best for well-defined domains such as mathematics problem solving (e.g., Koedinger & Corbett, 2006) and computer programming tasks (Corbett & Anderson, 1995), are used for inquiry practices or subskills that are similarly well-defined such as ‘identifying independent versus dependent variables’ in developing hypotheses or certain aspects of ‘data interpretation/analysis’.

By contrast, educational data mining techniques are used to assess inquiry practices for which skilled performance can manifest itself in several ways and more complex disambiguation of evidence is necessary; for a thorough review of the methods we used see Gobert et al., 2012, 2013; Sao Pedro et al., 2011. Two examples of inquiry practices that fall into this category are ‘testing stated hypotheses’ and ‘planning and carrying out experiments’ since there are a number of approaches that students can take on these, reflecting both skilled and unskilled performance (Shute, Glaser, & Raghavan, 1989; Kuhn, 2005).

We validated our assessment algorithms for these inquiry practices with thousands of middle school students. Specifically, our data-mined algorithm for evaluating whether students are testing their articulated hypothesis matches a human scorer 91% of the time. Similarly, our data-mined algorithm for assessing students’ skills at designing controlled experiments can distinguish controlled vs. confounded data collection 94% of the time (Sao Pedro et al., 2012). Furthermore, the latter algorithm can assess whether students are conducting controlled experiments even when students do not conduct their trials sequentially. Viewed this way, our work represents a large methodological advance over other assessments of this skill that evaluate only information from sequential trials as evidence of this skill (McElhaney & Linn, 2010; Klahr & Nigam, 2004) or evaluate only information from any two contrasting trials regardless of whether they are sequential or not. The former approach is too stringent an assessment of this since skill students do not necessarily conduct trials sequentially. The latter approach is too lenient an assessment of this skill since one cannot know whether the two non-sequential trials were conducted by chance or were collected deliberately to be contrasted by the student.

As of this writing, we have developed and validated assessment algorithms for all the NGSS science practices. We have also shown their generalizability across multiple domains (Sao Pedro, Jiang, Paquette, Baker, & Gobert, 2014; Sao Pedro, Gobert, Toto, & Paquette, 2015; Gobert, Kim, Sao Pedro, Kennedy, & Betts, in press).

Summary

In an educational system like the one in the United States, which is dedicated to standardization and accountability, inquiry in science classrooms cannot take center stage until considerable progress has been made on inquiry assessment. This is where "the rubber meets the road", and success with reform outlined in policy documents such as the NGSS rests, in our opinion, on rigorous assessment of students' science practices. In this chapter we described Inq-ITS, an on-line environment for assessing and scaffolding students' inquiry practices. We described how the NGSS, the learning sciences literature, the literature on students' difficulties with inquiry, and our early pilot work with students with hands-on inquiry tasks informed the design of Inq-ITS and how our system measured up in terms of granularity, measurability, and instructional relevance of the assessed scientific practices (Leighton & Gierl, 2011).

We argue that our system reflects several key advances in assessment design and practice in several areas. Perhaps most notably, the use of sophisticated knowledge-engineered and data-mined models allowed us to (1) assess students' inquiry practices in real time, (2) generate teacher reports and alerts in real time, and (3) trigger a pedagogical agent to scaffold inquiry in real time, all of which are critical to deep learning (Black & Wiliam, 1998; Pellegrino et al., 2001). This has several related benefits. First, teachers do not need to use additional instructional time for assessment because they receive immediate reports and alerts about their students and know who and what inquiry practices to focus on during instruction. Second, since our assessment algorithms work in real time over the web, the data-mined models are able to continually capture their emerging learning trajectory, which is ideal for continual adaptive assessment, adaptive instruction, and effective learning (Klahr & Nigam, 2004; Vygotsky, 1978).

Third, from a research perspective, the continual data that the Inq-ITS system provides allows us to advance our understanding of how students both conduct inquiry and hone these practices of inquiry over time. Finally, due to the sophistication of the microworld and widget design as well as the design of the underlying computational architecture, the Inq-ITS system can be scaled to many users simultaneously. Due to these functionalities, Inq-ITS and systems like it will continue to reduce or eliminate the separation between learning activities and assessment activities, allowing us to realize both the long-range vision for learner-centered environments (Quellmalz & Pellegrino, 2009; Quellmalz et al., 2012) as well as the rigorous assessment of inquiry practices as called for in the NGSS.

Janice Gobert (Ph.D. Cognitive Science, University of Toronto, 1994). Janice received her Ph.D. from the University of Toronto (OISE) in Applied Cognitive Science in 1994. She is a Professor of Learning Sciences and Educational Psychology at Rutgers University. Janice is also a Co-Founder and CEO of Apprendis. She has been Principal Investigator or Co-Principal Investigator on several grants to date totaling over $20M, all of which were focused on technology-based learning and assessment of learning in science. She also served as the N. American Editor of the International Journal of Science Education from 2001-2006.

Michael Sao Pedro (Ph.D. Learning Sciences; M.S., Computer Science, Worcester Polytechnic Institute). Michael did his Ph.D. under Janice Gobert's supervision while at Worcester Polytechnic Institute. He is a Co-Founder and the Chief Technology Officer of Apprendis. He specializes in the development of digital assessments for science using Educational Data Mining. Formerly, he was a Senior Software Engineer at BAE Systems (formerly ALPHATECH, Inc.). There, he led several artificial intelligence-inspired software efforts on several Phase I/II SBIR and DARPA projects.

References

Ahn, W., Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology, 31,* 82-123.

Ahn, W., Kalish, C., Medin, D., & Gelman, S. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54,* 299-352.

Aleven, V., & Koedinger, K. (2000). Limitations of Student Control: Do Students Know When They Need Help? In G. Gauthier, C. Frasson, & K. VanLehn (Ed.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000* (pp. 292-303). Berlin: Springer-Verlag.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward Tutoring Help Seeking. *Proceedings of the Seventh International Conference on Intelligent Tutoring Systems* (pp. 227-239). Berlin: Springer-Verlag.

American Association for the Advancement of Science. (1993*). Benchmarks for science literacy*: 1993. New York Oxford: Oxford University Press.

Amershi, S., 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., Corbett, A.T., Koedinger, K.R., Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167-207.

Aulls, M. W. (2002). The Contributions of Co-Occurring Forms of Classroom Discourse and Academic Activities to Curriculum Events and Instruction. *Journal of Educational Psychology, 94*(3), 520–538.

Baker, R., 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.

Behrens, J. (2009). Response to Assessment of Student Learning in Science Simulations and Games. White paper for Cisco Systems. (NB fuller ref needed).

Behrens, J. (2013). Harnessing the currents of the digital ocean. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Black, P., Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: School of Education, King's College.

Bransford, J., Brown, A. & Cocking, R. (2000). *How People Learn*. National Academy Press. Washington, D.C.

Brem, S., Rips, L. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24,* 573-604.

Brewer, W. F. (1987). Schemas versus mental models in human memory.

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. M. (Ed.), *Classroom lessons: integrating cognitive theory and classroom practice.* Cambridge, Massachusetts: MIT Press.

Buckley, B., Gobert, J. D., & Horwitz, P. (2006). Using log files to track students' model-based inquiry. *Proceedings of the 7th International Conference on Learning Sciences*, (pp. 57-63). Bloomington, IN.

Charney, D., Reder, L., & Kusbit, G.R. (1990) Goal setting and procedure selection in acquiring computer skills: A comparison of tutorials, problem solving, and learning exploration. *Cognition and Instruction*, 7(4), 323-342

Chen, Z., 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.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*(3), 271-315.

Chinn, C. A., 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.

Christensen, C., Horn, M. & Johnson, C. (2008). Disrupting class: How disruptive innovation will change the way the world learns. McGraw-Hill, NY, NY.

Clarke-Midura, J., Code, J., Zap, N. & Dede, C. (2012). Assessing science inquiry in the classroom: A case study of the virtual assessment project. In L. Lennex & K. Nettleton (Eds.), *Cases on Inquiry Through Instructional Technology in Math and Science: Systemic Approaches*. New York, NY: IGI Publishing.

Clarke-Midura, J., Dede, C. & Norton, J. (2011). The road ahead for state assessments. Policy Analysis for California Education and Rennie Center for Educational Research & Policy. MA: Rennie Center for Educational Research & Policy.

Clement, J. (1989). *Learning via model construction and criticism* (pp. 341-381). Springer US.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of research in science teaching*, *30*(10), 1241-1257.

Clement, J., Brown, B., & Zietsman, A. (1989). Not all preconceptions are misconceptions: Finding "anchoring conceptions" for grounding instruction on students' intuitions. *International Journal of Science Education,* 11, 554-565

connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, *55*(6), 503-513.

Corbett, A., Anderson, J. (1995). Knowledge-Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

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., Veenman, M., and Wilhelm, P. (2005). Determinants of discovery learning in a complex simulation learning environment. In P. Gardenfors and P. Johansson, (Eds.), *Cognition, Education, and Communication Technology,* Mawah, NJ: Erlbaum.

de Jong, T., van Joolingen, W.R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2), 179-201.

DeBoer, G., Abell C., Gogos, A., Michiels, A., Regan, T. & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. Project 2061. American Association for the Advancement of Science. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice*, NSTA Press.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. *Computers and their impact on state assessment: Recent history and predictions for the future*, 273-306.

Driver, R., & Erickson, G. (1983). Theories-in-Action: Some Theoretical and Empirical Issues in the Study of Students' Conceptual Frameworks in Science. Studies in Science Education, 10(1), 37-60. doi:10.1080/03057268308559904

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science: A Multidisciplinary Journal*, 17(3), 397-434.

Ericsson, K.A., Simon, H. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Feigenbaum, E.A., McCorduck, P. (1983). Land of the Rising Fifth Generation Computer. *High Technology*, 3(6), 64-70.

Frederiksen, C. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology, 7,* 471-458.

Frederiksen, C. (1986). *Cognitive models and discourse analysis*. In C. Cooper and S. Greenbaum (Eds.), Written communication annual volume 1: Linguistic approaches to the study of written discourse. Beverly Hills, CA: Sage.

Gentner, D., & Stevens, A. L. (1983). *Mental Models*, Erlbaum. Hillsdale, NJ.

Gilbert, J. K. (1993). *Models and modelling in science education.* Hatfield, Herts: Association for Science Education.

Gilbert, S. (1991). Model building and a definition of science. *Journal of Research in Science Teaching,* 28(1), 73-79.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. DeCorte, M. Linn, H. Mandl, & L. Verschaffel, *Computer-based Learning Environments and Problem-Solving* (pp. 345-371). Heidelberg, Germany: Springer-Verlag.

Gobert, J. (1994). *Expertise in the comprehension of architectural plans: Contribution of representation and domain knowledge.* Unpublished doctoral dissertation. University of Toronto, Toronto, Ontario.

Gobert, J. (in press). Microworlds. In Gunstone, R. (Ed.) *Encyclopedia of Science Education*. Springer.

Gobert, J. (2000). A typology of models for plate tectonics: Inferential power and barriers to understanding, *International Journal of Science Education*, 22, 937-977.

Gobert, J. (2005). Leveraging technology and cognitive theory on visualization to promote students' science learning and literacy. In *Visualization in Science Education*, J. Gilbert (Ed.), pp. 73-90. Springer-Verlag Publishers, Dordrecht, The Netherlands. ISBN 10-1-4020-3612-4.

Gobert, J. (in press). Microworlds. In Gunstone, R. (Ed.) *Encyclopedia of Science Education*. Springer.

Gobert, J., Buckley, B. (2000). Special issue editorial: Introduction to model-based teaching and learning. *International Journal of Science Education, 22(9),* 891-894.

Gobert, J., Clement, J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching,* 36(1), 39-53.

Gobert, J., Frederiksen, C. (1988). The comprehension of architectural plans by expert and sub-expert architects. *Proceedings of the Tenth Annual Meeting of the Cognitive Science Society.* Montreal, Canada, August 17-19, Hillsdale, NJ.: Lawrence Erlbaum. pp. 641-657.

Gobert, J.D., Kim, Y.J, Sao Pedro, M.A., Kennedy, M., & Betts, C.G. (in press). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity*. doi:10.1016/j.tsc.2015.04.008

Gobert, J., Sao Pedro, M., Baker, R.S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds, *Journal of Educational Data Mining, 15, Volume 4,* 153-185.

Gobert, J., Sao Pedro, M., Raziuddin, J. & Baker, R. (2013). From log files to assessment metrics for science inquiry using educational data mining. *Journal of the Learning Sciences, 22(4),* 521-563.

Gobert, J.D. (June, 2014). *Using Data Mining on Log files for Real Time Assessment and Learning*. Featured speaker at the Cyberlearning Summit, June 9-10, Madison, WI.

Gobert, J.D., Baker, R., & Wixon, M. (2015). Operationalizing and Detecting Disengagement Within On-Line Science Microworlds. *Educational Psychologist, 50(1), 43-57*. DOI:10.1080/00461520.2014.999919

Haertel, G., Lash, A., Javitz, H., & Quellmalz, E. (2006). *An instructional sensitivity study of science inquiry items from three large-scale science examinations*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hanushek, E. A. (2005). The economics of school quality. *German Economic Review*, *6*(3), 269-286.

Harrison, A.M., Schunn, C.D. (2004). The transfer of logically general scientific reasoning skills. In Forbus K., Gentner, D., and Regier, T. (Eds.) *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pp. 541–546. Mahwah, NJ: Erlbaum.

Hmelo-Silver, C. E., Duncan, R. G., & 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.

Huff, K. & Goodman, D. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton, & M. J. Gierl (Eds). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.

Keselman, A. (2003). Promoting scientific reasoning in a computer-assisted environment. *Journal for Research in Science Teaching, 40,* 898–921.

Ketelhut, D. J., & Dede, C. (2006). Assessing Inquiry Learning. Paper presented at the National Association for Research in Science Teaching, April 3-6, San Francisco, CA.

Kindfield, A.C.H. (1993). Biology Diagrams: Tools to think with. *Journal of the Learning Sciences,* 3(1), 1-36.

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction dose not work: An analysis of the failure of constructivist, discover, problem-based, experimental, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86.

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

Klahr, D., Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science, 15(10)*, 661-667.

Klayman, J., Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211-228.

Koedinger, K. R., & Aleven V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19(3),* 239-264.

Koedinger, K. R., Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.) *The Cambridge Handbook of the Learning Sciences,* (pp. 61-78). Cambridge University Press.

Krajcik, J.S., Blumenfeld, P., Marx, R.W., Bass, K.M., Fredricks, J., & Soloway, E. (1998). Middle school students' initial attempts at inquiry in project-based science classrooms. *Journal of the Learning Sciences. 7(3&4)*, 313-350.

Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000). *Reforming Science Education through University and School District Collaborations*. Conference paper presented at NARST.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96,* 674-689.

Kuhn, D. (1991). *The skills of argument.* Cambridge: Cambridge University Press.

Kuhn, D. (2005). *Education for thinking.* Cambridge, MA: Harvard University Press.

Kuhn, D., Dean, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5,* 261-288.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Society for Research in Child Development Monographs 60(4, Serial No. 245)*.

Kuhn, D., Katz, J., & Dean, D. (2004). Developing reason. *Thinking and Reasoning, 10,* 197-219.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning, *Cognition and Instruction, 9*(4), 285-327.

Larkin, J. & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-100.

Leeman-Munk, S., Wiebe, E., & Lester, J. (2013). Mining Student Science Argumentation Text to Inform an Intelligent Tutoring System. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Leighton, J. & Gierl, M. (2011). The learning science in educational assessment: The role of cognitive models. Cambridge University Press, NY, NY.

Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*.

Lowe, R.K. (1989). Search strategies and inference in the exploration of scientific diagrams. *Educational Psychology*, 9(1), 27-44.

McDermott, L., Rosenquist, M. L. & van Zee, E. H. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics, 55,* 503-513.

McElhaney, K., Linn, M. (2008). Impacts of Students' Experimentation Using a Dynamic Visualization on their Understanding of Motion. *Proceedings of the 8th International Conference of the Learning Sciences, ICLS 2008, Volume 2* (pp. 51-58). Ultrecht, The Netherlands: International Society of the Learning Sciences, Inc.

McElhaney, K., Linn, M. (2010). Helping Students Make Controlled Experiments More Informative. In K. Gomez, L. Lyons, & J. Radinsky (Ed.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010) - Volume 1, Full Papers* (pp. 786-793). Chicago, IL: International Society of the Learning Sciences.

McNeill, K. L., Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In Lovett, M & Shah, P (Eds.), *Thinking with data.* (pp. 233-265). New York, NY: Taylor & Francis Group, LLC.

Mislevy, R. Behrens, J., Dicerbo, K., & Levy, R. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4*(1), 49-110.

Mislevy, R., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., & Haertel, G. (2003). *Design patterns for assessing science inquiry.* Menlo Park, CA: SRI International.

Mitrovic, A., Mayo, M., Suraweera, P., & Martin, B. (2001). Constraint-based tutors: A success story. In L. Monostori, J. Vancza, & M. Ali (Ed.), *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*, pp. 931-940. Budapest, Hungary: Springer-Verlag.

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS.

National Research Council (2007). Taking Science to School: Learning and Teaching Science in Grades K-8. R. Duschl, Heidi Schweingruber, & A. Shouse (Eds.) Washington, DC: The National Academies Press.

National Research Council. (1996). *National Science Education Standards*, Washington, D.C.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A. and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ, Prentice Hall.

Njoo, M., 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.

Organization for Economic Cooperation and Development (2014). *PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know.*

Papert, S. (1980). Computer-based microworlds as incubators for powerful ideas. In R. Taylor, *The Computer in the School: Tutor, Tool, Tutee* (pp. 203-201). New York, NY: Teacher's College Press.

Papert, S. (1993). *Mindstorms: children, comptuers, and powerful ideas, 2nd Edition.* New York: Basic Books.

Pea, R. D., & Kurland, D. M. (1984). On the cognitive effects of learning computer programming. *New ideas in psychology*, *2*(2), 137-168.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Perkins, D. (1986). *Knowledge as design*. Hillsdale, NJ: Erlbaum.

Pfundt, H., & Duit, R. (1994). Bibliography, students' alternative frameworks and science

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*. Information Age.

Quellmalz, E. S., Timms, M., & Buckley, B. (2009). *Using science simulations to support powerful formative assessments of complex science learning*. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Quellmalz, E., Haertel, G. (2004). Technology supports for state science assessment systems *Paper commissioned by the National Research Council Committee on Test Design for K–12 Science Achievement*. Washington, DC: National Research Council.

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., & 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.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Il.

Quellmalz, E., Pellegrino, J. (2009). Technology and Testing. *Science, 323*, 75-79.

Quellmalz, E., Timms, M., & Schneider, S., (2009). *Assessment of Student Learning in Science Simulations and Games*. Washington, DC: National Research Council Report, Washington, DC.

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.

Raghavan, K., Glaser, R. (1995). Model-based analysis and reasoning in science: The MARS curriculum. *Science Education, 79,* 37-61.

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

Resnick, L. B. (1997). Getting to work: Thoughts on the function and form of the school-to-work transition. *Papers and Proceedings. Transitions in work and learning: Implications for assessment*, 249-263.

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

Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135-146.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, *100*(3), 349.

Rupp, A.A., Gushta, M., Mislevy, R.J., 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). Available online at http://escholarship.bc.edu/jtla/vol8/4.

Sadler, T. D. (2004). Informal reasoning regarding socio-scientific issues: A critical review of research. *Journal of research in science teaching*, *41*(5), 513-536.

Sao Pedro, M., Jiang, Y., Paquette, L., Baker, R.S. & Gobert, J. (2014). Identifying Transfer of Inquiry Skills across Physical Science Simulations using Educational Data Mining. In *Proceedings of the 11th International Conference of the Learning Sciences*. Boulder, CO (pp. 222-229)

Sao Pedro, M, Gobert, J.D., & Betts, C.G. (2015). *Inq-Blotter: Revolutionizing How Teachers Identify and Support Students Needing Help During Inquiry.* Proposal (EDIES15C0018) funded by the U.S. Department of Education SBIR program.

Sao Pedro, M. A., Baker, R. S., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. In R. Baker, A. Merceron, & P. Pavlik (Ed.), *Proceedings of the 3rd International Conference on Educational Data Mining*, (pp. 181-190). Pittsburgh, PA.

Sao Pedro, M., Gobert, J., Toto, E., & Paquette, L. (April, 2015). *Assessing Transfer of Students' Data Analysis Skills across Physical Science Simulations*. Paper presented as part of Bejar, I. et al.'s symposium on The State of the Art in Automated Scoring of Science Inquiry Tasks at the Annual Meeting of the American Education Research Association. Chicago, IL.

Sao Pedro, M. A., Gobert, J. D., & Raziuddin, J. (2010). Comparing Pedagogical Approaches for the Acquisition and Long-Term Robustness of the Control of Variables Strategy. In K. Gomez, L. Lyons, & J. Radinsky (Ed.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences, ICLS 2010, Volume 1, Full Papers* (pp. 1024-1031). Chicago, IL: International Society of the Learning Sciences.

Sao Pedro, M., Baker, R., & Gobert, J. (2012). Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. In *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP 2012)*. Montreal, QC, Canada (pp. 249-260).

Sao Pedro, M**.**, Baker, R., & Gobert, J. (2013b). What Different Kinds of Stratification Can Reveal about the Generalizability of Data-Mined Skill Assessment Models. In *Proceedings of the 3rd Conference on Learning Analytics and Knowledge.* Leuven, Belgium.

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2011). Using Machine-Learned Detectors of Systematic Inquiry Behavior to Predict Gains in Inquiry Skills. *User Modeling and User-Adapted Interaction*. DOI: 10.1007/s11257-011-9101-0

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2013a). Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction, 23,* 1-39.

Sao Pedro, M., Gobert, J., & Baker, R. (2012). Assessing the Learning and Transfer of Data Collection Inquiry Skills Using Educational Data Mining on Students' Log Files. *Paper presented at The Annual Meeting of the American Educational Research Association.* , (pp. Retrieved April 15, 2012, from the AERA Online Paper Repository). Vancouver, BC, Canada.

Sao Pedro, M.A. (2013). *Real-time Assessment, Prediction, and Scaffolding of Middle School Students' Data Collection Skills within Physical Science Simulations*. Social Science and Policy Studies: Learning Sciences and Technologies Program Ph.D. Dissertation. Worcester Polytechnic Institute Technical Report etd-042513-062949.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31-57.

Schauble, L., Glaser, R., Duschl, R., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences, 4*(2), 131-166.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *Journal of the Learning Sciences*, 1 (2), 201-238.

Schauble, L., Klopfer, L.E. and Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28* (9), 859-882.

Schunn, C. D., Anderson, J. R. (1998). Scientific Discovery. In J. R. Anderson, *The Atomic Components of Thought* (pp. 385-428). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V., Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 55-71.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In *Learning and Individual Differences: Advances in Theory and Research.* (pp. 279–326). New York, NY: W.H. Freeman

Timms, M., Clements, D.H., Gobert, J., Ketelhut, D.J., Lester, J., Reese, D.D., & Wiebe, E. (2012). New measurement paradigms. A report prepared for the Community for Advancing Discovery Research in Education (CADRE), United States. Available at: http://works.bepress.com/michael_timms/36

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51(1), 1–10.

van Joolingen, W.R., & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science*, 25(5), 307-346.

van Joolingen, W.R., de Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science*, 20(5-6), 389-404.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

White, B. & Frederiksen, J. (1990). Causal model progressions as a foundation for intelligent learning environments. *Artificial Intelligence*, 24, 99-157.

White, B.Y. (1993). Intermediate causal models: A missing link for successful science education. In R. Glaser *Advances in Instructional Psychology.* (Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers) 177-252

Wiggins, G. (1998). *Educative assessment: designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Table 1

Next Generation Science Standards (NGSS Lead States, 2013)

Life Science	**Physical Science**
LS1: From Molecules to Organisms: Structures and Processes LS2: Ecosystems: Interactions, Energy, and Dynamics LS3: Heredity: Inheritance and Variation of Traits LS4: Biological Evolution: Unity and Diversity	PS1: Matter and Its Interactions PS2: Motion and Stability: Forces and Interactions PS3: Energy PS4: Waves and Their Applications in Technologies for Information Transfer
Earth & Space Science	**Engineering & Technology**
ESS1: Earth's Place in the Universe ESS2: Earth's Systems ESS3: Earth and Human Activity	ETS1: Engineering Design ETS2: Links Among Engineering, Technology, Science, and Society

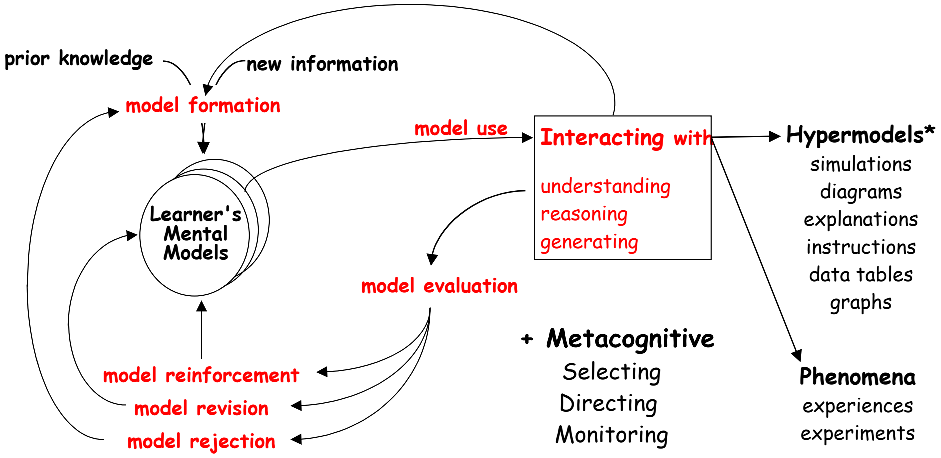
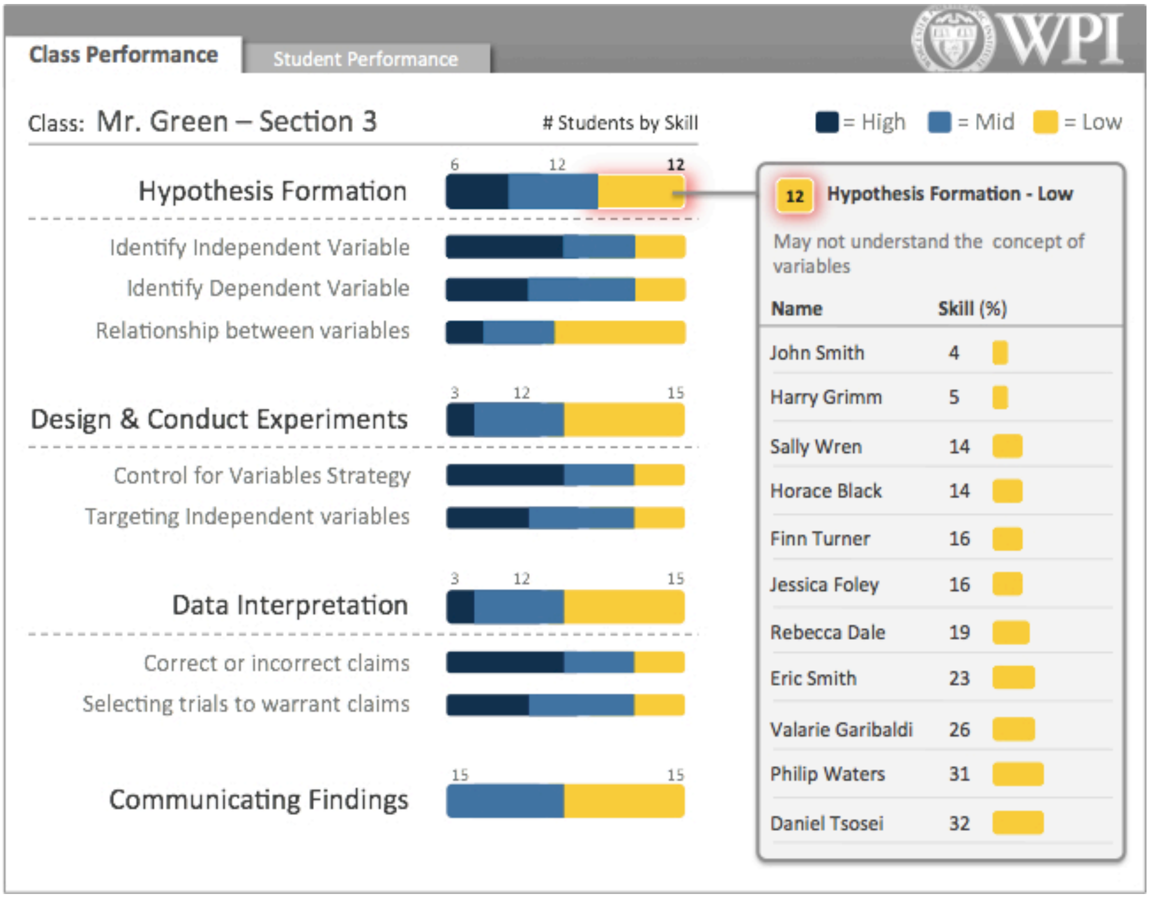


Figure 1 Model-based learning and teaching framework.

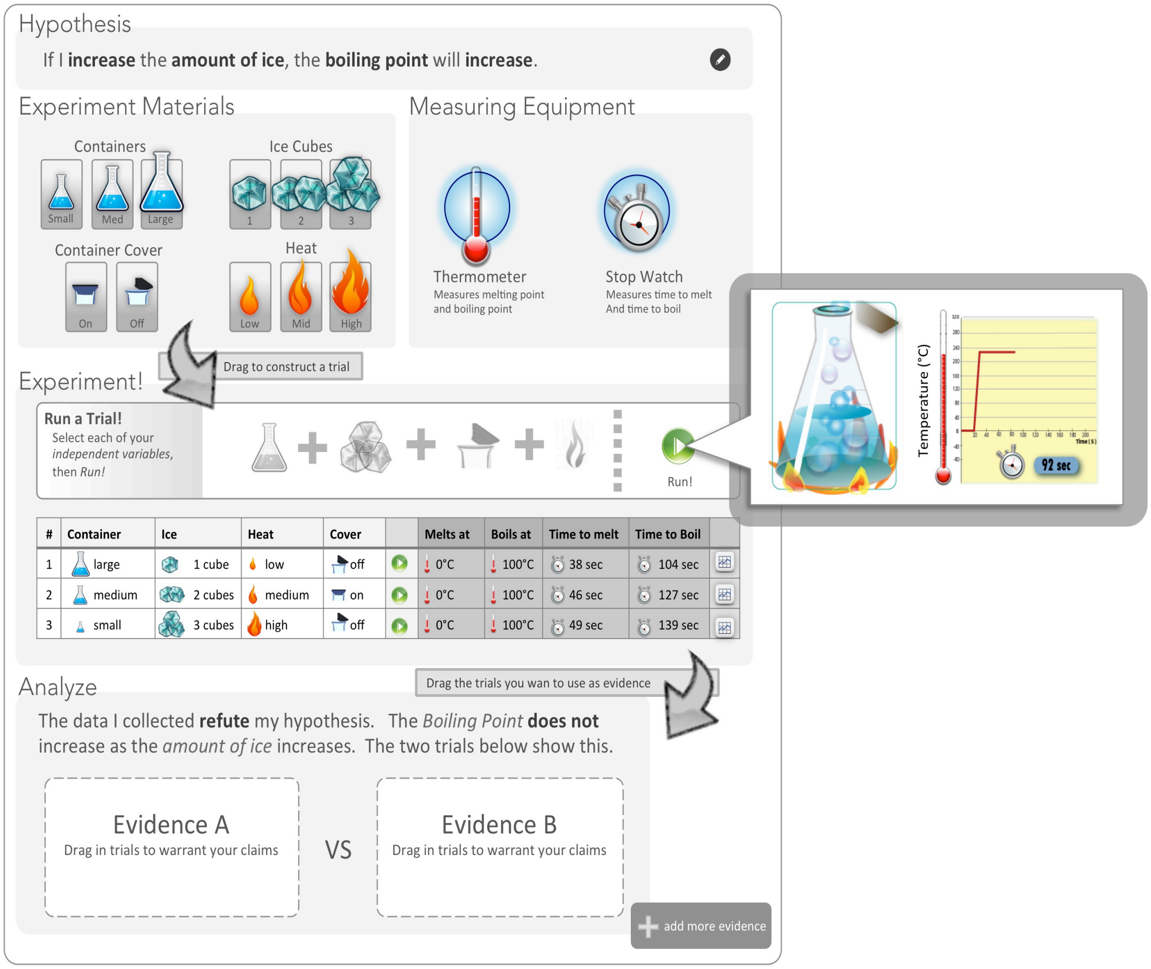

WPI
Class Performance
Student Performance
Class: Mr. Green – Section 3
# Students by Skill
■ = High
■ = Mid
■ = Low
Hypothesis Formation
6
12
12
Identify Independent Variable
Identify Dependent Variable
Relationship between variables
Design & Conduct Experiments
3
12
15
Control for Variables Strategy
Targeting Independent variables
Data Interpretation
3
12
15
Correct or incorrect claims
Selecting trials to warrant claims
Communicating Findings
15
15

12 Hypothesis Formation - Low

May not understand the concept of variables

Name	Skill (%)
John Smith	4
Harry Grimm	5
Sally Wren	14
Horace Black	14
Finn Turner	16
Jessica Foley	16
Rebecca Dale	19
Eric Smith	23
Valarie Garibaldi	26
Philip Waters	31
Daniel Tsosei	32

Figure 2 Assessment report for teachers reported out by inquiry practice and sub-skills.



Hypothesis

If I **increase** the **amount of ice**, the **boiling point** will **increase**.

Run a Trial!
Select each of your *independent variables*, then *Run!*

#	Container	Ice	Heat	Cover		Melts at	Boils at	Time to melt	Time to Boil	
1	large	1 cube	low	off	▶	0°C	100°C	38 sec	104 sec	
2	medium	2 cubes	medium	on	▶	0°C	100°C	46 sec	127 sec	
3	small	3 cubes	high	off	▶	0°C	100°C	49 sec	139 sec	

The data I collected **refute** my hypothesis. The *Boiling Point* **does not** increase as the *amount of ice* increases. The two trials below show this.

Evidence A
Drag in trials to warrant your claims
VS
Evidence B
Drag in trials to warrant your claims
add more evidence

Figure 3 Screenshot of the Inq-ITS system with all phases of inquiry shown. Students generate a question, run trials to test it, then interpret data, and select trials to warrant claims with evidence.

Acknowledgements
This research is funded by the Department of Education (R305A120778). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies.