Journal of the Learning Sciences

ISSN: 1050-8406 (Print) 1532-7809 (Online) Journal homepage: https://www.tandfonline.com/loi/hlns20

From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining

Janice D. Gobert , Michael Sao Pedro , Juelaila Raziuddin & Ryan S. Baker

To cite this article: Janice D. Gobert , Michael Sao Pedro , Juelaila Raziuddin & Ryan S. Baker (2013) From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining, Journal of the Learning Sciences, 22:4, 521-563, DOI: 10.1080/10508406.2013.837391

To link to this article: https://doi.org/10.1080/10508406.2013.837391

THE JOURNAL OF THE LEARNING SCIENCES, 22: 521–563, 2013

ISSN: 1050-8406 print / 1532-7809 online
DOI: 10.1080/10508406.2013.837391

Routledge
Taylor & Francis Group

From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining

Janice D. Gobert and Michael Sao Pedro
Learning Sciences & Technologies Graduate Program
Worcester Polytechnic Institute
Apprendis, LLC

Juelaila Raziuddin
Learning Sciences & Technologies Graduate Program
Worcester Polytechnic Institute

Ryan S. Baker
Department of Human Development
Teachers College Columbia University

We present a method for assessing science inquiry performance, specifically for the inquiry skill of designing and conducting experiments, using educational data mining on students' log data from online microworlds in the Inq-ITS system (Inquiry Intelligent Tutoring System; www.inq-its.org). In our approach, we use a 2-step process: First we use text replay tagging, a type of rapid protocol analysis in which categories are developed and, in turn, used to hand-score students' log data. In the second step, educational data mining is conducted using a combination of the text replay data and machine-distilled features of student interactions in order to produce an automated means of assessing the inquiry skill in question; this is referred to as a *detector.* Once this detector is appropriately validated, it can be applied to students' log files for auto-assessment and, in the future, to drive scaffolding in real time. Furthermore, we present evidence that this detector developed in 1 scientific domain, phase change, can be used—with no modification or retraining—to effectively detect science inquiry skill in another scientific domain, density.

Correspondence should be addressed to Janice D. Gobert, Learning Sciences & Technologies Graduate Program, Department of Social Sciences & Policy Studies, Salisbury Labs, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. E-mail: jgobert@wpi.edu

OVERVIEW

In the present article, we describe how science inquiry skills can be assessed within the Inq-ITS system (Inquiry Intelligent Tutoring System, formerly referred to as *Science Assistments;* Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012) by leveraging techniques from educational data mining (Baker & Yacef, 2009; Romero & Ventura, 2010). Our system aims to enable automated, authentic, scalable assessment of middle school students' inquiry skills across several topics in physical, life, and earth science using interactive simulations. It also aims, in time, to provide real-time, automated support and feedback as students conduct their inquiry investigations. Our activities serve as performance assessments of inquiry skills; the actions students take within the simulation are the bases for assessment (Gobert et al., 2012). Here we focus on our work toward developing an automated assessment referred to as a *detector* that assesses one key inquiry skill, designing controlled experiments (National Research Council [NRC], 1996), as students conduct investigations in one of our physical science simulations.

In our approach, we use a two-step process that leverages human judgment for this inquiry skill in order to develop a detector that replicates human judgment. In the first step, we hand-score students' log data on the inquiry skill, designing controlled experiments. In the second step, data mining is conducted to develop models that can predict human judgments from features of students' interactions in the microworld. This results in a detector that can identify whether students demonstrate the inquiry skill of interest on a particular inquiry task or subtask. Once this detector is appropriately validated, it is applied to students' log files to automatically assess the skill. In the future, this detector will be used to drive scaffolding in real time as well.

This article has four key goals. First, we briefly present a history of inquiry, its definition, and it treatment in the science standards; second, we review criticisms of current approaches to the assessment of inquiry. Third, we describe how our detector for one inquiry skill of interest, designing controlled experiments, was developed and validated (Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013; Sao Pedro, Gobert & Baker, 2012). Fourth, we present evidence that this detector is generalizable by showing that it can be used to assess whether students demonstrate this skill in a *different* science topic than that in which the detector was originally built; that is, we demonstrate that our detector developed in phase change can be used—with no modification or retraining to the machine-learned algorithms—to effectively assess science inquiry skill in another domain, density. Together, the article provides a framework and technical solution for how to use educational data mining (Baker & Yacef, 2009; Romero & Ventura, 2010) to develop a rigorous method of assessing scientific inquiry skills in real time that is also scalable. We argue that this approach makes a significant contribution to

the auto-scoring of ill-defined skills as most work to date has been conducted in well-defined domains such as math (Koedinger & Corbett, 2006; Razzaq et al., 2005).

BACKGROUND LITERATURE

The Origins of Inquiry in Education

John Dewey, a former science teacher, first introduced the idea of inquiry in education in 1910. He claimed that there was too heavy an emphasis on rote science facts in science instruction. According to Dewey (1910), learning science methods should consist of "sensing perplexing situations, clarifying the problem, formulating a tentative hypothesis, testing it, revising with rigorous tests, and acting on the solution" (pp. 121–127). Dewey later modified this earlier interpretation and added presentation of the problem, formulation of a hypothesis, collection of data during the experiment, and formulation of a conclusion (Dewey, 1916).

Following Dewey and the launching of Sputnik in October 1957, policymakers in the United States began to question the quality of science instruction in schools, which, in turn, instantiated a call for change in all science curricula. The National Science Foundation provided grants for the creation of, implementation of, and teacher professional development of new science curricula with an emphasis on "thinking like a scientist" and science inquiry skills such as observing, classifying, inferring, and controlling variables (as discussed in DeBoer, 1991). Following this, the American Association for the Advancement of Science delineated what all American students should know by the end of 12th grade in *Science for All Americans* (Rutherford & Ahlgren, 1989), describing inquiry as a content topic. Policy documents subsequently put out by the NRC (1996, 2000) described inquiry as the overarching goal of scientific literacy. In brief, the NRC (2000) summarized the five essential features of inquiry, regardless of grade level, as including the following:

> 1. Scientifically-oriented questions that engage students; 2. evidence collected by students that allows them to develop and evaluate their explanations to the scientifically oriented questions; 3. explanations developed by students from their evidence to address the scientifically oriented questions; 4. evaluation of their explanations, which can include alternative explanations that reflect scientific understanding; and 5. communication and justification of their proposed explanations. (p. 29)

The level of sophistication of each of these types of tasks varies by grade level, with increasing complexity at the upper grade levels (NRC, 2000).

Despite the well-acknowledged need in these policy documents (NRC, 1996, 2000; Rutherford & Ahlgren, 1989) for inquiry skills to be considered as an integral component of science literacy, there was growing concern on the part of policymakers that the state assessments used to assess students' science knowledge were not well suited to assessing science process skills (DeBoer et al., 2008). In brief, these standardized achievement tests for science assessed proficiency with pencil-and-paper multiple-choice items and were thought to perhaps be better designed to capturing rote understanding of science than inquiry skills. More recently, there has also been an issue as to whether they align well with current national standards (Quellmalz, Kreikemeier, DeBarger, & Haertel, 2007). Furthermore, because they are product rather than process oriented, they are not capable of providing rich data about students' inquiry strategies (Gobert et al., 2012), nor are they likely to elicit inquiry-related reasoning from students (Clarke-Midura, Dede, & Norton, 2011). In terms of measurement, these tests may not adequately identify inquiry skills (Black, 1999; Pellegrino, Chudowsky, & Glaser, 2001), and they do not provide precise and replicable measures of such skills (Clarke-Midura et al., 2011; DeBoer et al., 2008; Haertel, Lash, Javitz, & Quellmalz, 2006; Quellmalz & Haertel, 2004; Quellmalz et al., 2007; Wilson & Bertenthal, 2006). Additional measurement difficulties stem from the large amount of data required for reliable assessment (Shavelson, Wiley, & Ruiz-Primo, 1999) and the difficulty of separating inquiry from the domain-specific context in which it was learned (Mislevy et al., 2003). Lastly, these types of assessments only permit feedback typically weeks or months after the assessment has been completed, which is too late for effective corrective intervention by the teacher. DiCerbo and Behrens (2012) described these limitations as an artifact of two related issues: a simplified conceptualization of both (a) the nature of the content or skill(s) being measured and (b) the type of data needed for the type of knowledge being assessed. In brief, the tests reflect, at least in part, the ways in which the knowledge or the skills being assessed were conceptualized at the time when the tests were designed (Mislevy, Behrens, DiCerbo, & Levy, 2012). Thus, given that this simplified view of science that does not take into account the richness of science process skills, it is understandable that standardized science assessments are not rich enough to capture and evaluate students' skills at conducting inquiry.

Toward Performance Assessments of Inquiry Skills

Hands-on performance assessments of inquiry, as an alternative to standardized tests, have been developed to measure higher level skills such as problem-solving for science (Linn, 1994), but they have been largely disregarded as useful for wide-scale use (Clarke-Midura, Code, Zap, & Dede, 2012). Specifically, studies on their reliability, construct validity, and scalability (Linn, 2000) have revealed problems

(Clarke-Midura et al., 2012) compared to multiple-choice tests (Linn, Baker, & Dunbar, 1991). In addition, despite their intuitive appeal, they have turned out to be cost prohibitive (Stecher & Klein, 1997), and their highly structured and short format (15–40 min) limits the investigation strategies that can be measured (Quellmalz et al., 2007).

In recent years, computer-based environments that include *microworlds,* computer-based environments that enable the study of scientific phenomena in a dynamic, interactive way (Gobert, in press; Papert, 1980), are providing new possibilities for assessing science and are being considered as alternatives to traditional assessments of inquiry (Behrens, 2013; Gobert et al., 2012; Pellegrino et al., 2001; Quellmalz, Timms, & Schneider, 2009). Key organizations and their respective policy documents, such as the National Assessment of Educational Progress (National Assessment Governing Board, 2011), Programme for International Student Assessment (PISA; Organisation for Economic Cooperation and Development, 2010), the National Educational Technology Plan, and NRC (2011), all acknowledge the benefits and potential of technology-based systems for the assessment of science inquiry.

These environments, which generate logs of students' actions in a nonintrusive way (Pellegrino et al., 2001) can provide a more fertile basis upon which to generate performance-based assessments of rich inquiry processes because student inquiry processes within these environments are captured in situ (Clarke-Midura et al., 2011). In addition, because log files are generated as students conduct inquiry, it is possible to assess students' inquiry processes in addition to the artifacts or products they create as a result of those processes (cf. Rupp, Gushta, Mislevy, & Shaffer, 2010).

In our learning environment (www.inq-its.org, formerly known as *Science Assistments;* Gobert et al., 2012), we provide students with microworlds for inquiry. Commensurate with Kuhn's conceptualization of science inquiry, "students investigate a set of phenomena—virtual or real—and draw conclusions about the phenomena" (Kuhn, Black, Keselman, & Kaplan, 2000, p. 497). Microworlds, like those embedded in our learning environment, afford authenticity because they share many features with real apparati for "doing science" (Gobert, 2005). Specifically, microworlds provide authentic performance assessment opportunities for inquiry skills because by using a microworld, students can generate a hypothesis, test it, interpret data, warrant claims with data, and communicate findings with regard to what they discover as they explore. These are the inquiry tasks reflected in the national frameworks (NRC, 1996, 2000), and they arguably represent the key skills used by scientists. Furthermore, when students conduct several tasks within a microworld or across multiple microworlds, this provides a large number of inquiry assessment opportunities, facilitating reliable assessment (Shavelson et al., 1999) in the context in which the skills are developing (Fadel, Honey, & Pasnick, 2007; Mislevy et al., 2003).

Critical Issues to Consider From a Measurement Perspective in Using Virtual Environments for Assessing Science Inquiry Skills

Despite the many affordances that microworlds provide for assessment, research has shown that these learning environments are not yet used for performance assessment to a substantial degree, except in a small number of projects (described later). For example, even in projects that use microworlds to foster learning, conventional paper-and-pencil assessments are typically used as posttests to assess learning (see the meta-analysis in Scalise, Timms, Clark, & Moorjani, 2009). In brief, despite the existence of rich authentic environments, these environments are not being used to measure the complex science knowledge and skills that they were designed to foster (Quellmalz et al., 2009). There are several possible reasons for this.

The principal issue has to do with the complexity of the resultant log data from learning environments, which makes it difficult to measure inquiry using conventional methods (Quellmalz et al., 2009). Specifically, these environments have three characteristics (Williamson, Mislevy, & Bejar, 2006) that add considerable complexity to inquiry skills assessment. First, the tasks include many domain-relevant steps/subtasks that are much more complex than typical multiple-choice items in terms of students' choices, actions, an so on. Second, there is potential for wider variability in the data for each of these tasks and subtasks that are captured in students' log files (than is seen in multiple-choice items), and this needs to be considered in the determination of skills. Third, the inquiry tasks and subtasks completed as part of an inquiry cycle are not independent from one another, and thus the assumptions of conditional independence (made in classical test theory) do not hold (Mislevy et al., 2012).

Another problem to using log data for performance assessment is that there is a lack of theory-based principles for distilling, parsing, and aggregating the large volumes of log data that are generated as students work in these environments (Gobert et al., 2012; Quellmalz et al., 2009) so that data can be interpreted in relation to pedagogical theories and in instructionally meaningful ways. For these reasons, traditional statistical measurement techniques designed to handle a set of summative data gathered via typical multiple-choice items designed to be independent from one another will not suffice as techniques for analyzing log data.

Key Projects in the Assessment of Inquiry Using Simulations and Virtual Worlds

There is some notable work on performance assessment with a wide range of analytical methods but whose goals are similar to ours. We review these here. A fuller description of many of these projects and their methodological approaches can be found in Timms et al. (2012).

Edys Quellmalz and her colleagues have a longstanding program of research on science assessment. In the Calipers I and Calipers II projects (www.calipers.sri.com), students use rich simulations to investigate scientific phenomena; students use tools such as an arrow tool to summarize the relationships between organisms in an ecosystem. These data are triangulated with summative multiple-choice items and are analyzed using a combination of methods such as multidimensional item response theory and confirmatory factor analysis to develop assessments from which students' inquiry skills are inferred (Quellmalz et al., 2008). In a newer project called Sim Scientists, Quellmalz and her colleagues (Quellmalz, Silberglitt, & Timms, 2011; Quellmalz, Timms, Buckley, et al., 2012; Quellmalz, Timms, Silberglitt, & Buckley, 2012) promote students' inquiry using a model-based teaching and learning framework (cf. Gobert & Buckley, 2000) and measure students' inquiry skills within the learning environment. These assessments, although primarily formative in nature, represent a richer approach to inquiry assessment than traditional multiple-choice tests.

Chris Dede and his former students have made considerable progress on performance assessment for inquiry skills. Their set of projects uses virtual environments in which an avatar, as an agent of the student, conducts inquiry. In the River City project (Ketelhut & Dede, 2006), students work in teams to engage in many inquiry tasks. Traditional pre/post assessments of reasoning in addition to a "letter to the mayor" task are used to assess students' learning. Similarly, in the Virtual Performance Assessment project (Clarke-Midura et al., 2012), which also uses River City as its platform, researchers measure students' skills at hypothesis generation, data collection, experimental design, and formulation of scientific explanations. In these projects, techniques such as item response theory and Bayesian knowledge tracing are used to assess students' skills. In SAVE Science (Situated Assessment using Virtual Environments for Science Content and Inquiry; Ketelhut, Nelson, Sil, & Yates, 2013), middle school students conduct inquiry in an environment called Scientopolis; the open response and log data are used to assess students' skills at analyzing and interpreting data. Automatic assessment is done using support vector machines (Vapnik, 1995) for predicting students' understanding of inquiry on summative open response questions.

James Lester and his colleagues (cf. Rowe & Lester, 2010), working in the context of a three-dimensional immersive environment on microbiology called Crystal Island, developed dynamic Bayesian network models of middle school students' narrative, strategic, and curricular knowledge. In more recent work by this group on the Leonardo project, the goal is to develop automated methods of assessing students' skills at formulating explanations using evidence (Leeman-Munk, Wiebe, & Lester, 2013). Like Ketelhut et al. (2013), the group is using support vector machines (Vapnik, 1995) to train a model to automatically score students' typed responses.

Also important to note are the assessment items by the National Assessment of Educational Progress and the American Association for the Advancement of Science. In 2009, the National Assessment of Educational Progress used interactive computer tasks with the goal of assessing high-level skills such as searching a database, selecting relevant information, and applying it to a problem. However, although these are important skills, they are not inquiry skills per se. In addition, these items are not yet well studied from a measurement point of view (Fu, Raizen, & Shavelson, 2009). The American Association for the Advancement of Science, although acknowledging the limitations of multiple-choice assessments, has developed assessment items about simple contexts with which students should be familiar. These items are designed to tap students' knowledge of the control of variables strategy (CVS; Chen & Klahr, 1999) and attempt to avoid confounding of context with CVS by designing contexts based on prior science content learning (G. DeBoer, personal communication, December 27, 2012). However, these assessments use a multiple-choice format; thus, they cannot assess students' inquiry processes. Lastly, PISA 2006 (Organisation for Economic Co-operation and Development, 2010) developed and conducted a pilot test of computer-based assessments for science to address one of their key components of scientific inquiry, namely, science competencies—students' skills at identifying scientific issues, explaining science phenomena, and using scientific evidence (Bybee, McCrae, & Laurie, 2009). These assessments were considered an advance over former versions of tests of science inquiry skills because they were aimed at skills not assessed in paper-based booklets (Koomen, 2006).

Prior Work on Designing and Conducting Experiments

Many studies have shown that students have difficulty designing and conducting experiments. In brief, students show a bias toward investigating features believed to be causal as opposed to noncausal (Kuhn, Schauble, & Garcia-Mila, 1992). Moreover, they may gather insufficient evidence to test hypotheses (Schauble, Glaser, Raghavan, & Reiner, 1991; Shute & Glaser, 1990), for example by running only one trial (Kuhn et al., 1992), which makes it impossible to find out how an independent variable influences a dependent variable. Students also run the same trial repeatedly (Buckley, Gobert, & Horwitz, 2006; Kuhn et al., 1992), change too many variables in the same trial (Glaser, Schauble, Raghavan, & Zeitz, 1992; Harrison & Schunn, 2004; Kuhn, 2005a; McElhaney & Linn, 2008, 2010; Reimann, 1991; Schunn & Anderson, 1998, 1999; Shute & Glaser, 1990; Tsirgi, 1980), and do not use the range of potential informative experiments available (e.g., only testing two levels out of three of an independent variable; Harrison & Schunn, 2004; Kuhn et al., 1992; Schunn & Anderson, 1999). They also run experiments that try to achieve an outcome (e.g., make something burn as quickly as possible) as opposed to testing a hypothesis (Njoo & de Jong, 1993; Schauble, Glaser, Duschl, Schulze, & John, 1995; Schauble, Klopfer, & Raghavan, 1991).

Other prior studies relevant to the present work have addressed principally how to best teach, rather than assess, CVS. For example, several studies have addressed the efficacy of strategies for teaching or fostering CVS (e.g., Chen & Klahr, 1999; Ford, 2012; Klahr & Nigam, 2004; Sao Pedro, Gobert, Heffernan, & Beck, 2009; Sao Pedro, Gobert, & Raziuddin, 2010; Siler, Klahr, Magaro, Willows, & Mowery, 2010; Strand-Cary & Klahr, 2008; Zohar & David, 2008). Some studies with goals more similar to ours have analyzed performance at designing controlled experiments during open-ended inquiry tasks involving microworlds (e.g., Kuhn & Pease, 2008; McElhaney & Linn, 2008, 2010; Schunn & Anderson, 1998; Shute & Glaser, 1990). Of those, Shute and Glaser (1990), Schunn and Anderson (1998), and McElhaney and Linn (2008, 2010) developed scoring techniques for determining whether students were designing controlled experiments by examining the number of *successive* pairwise CVS trials found in students' log files.

In McElhaney and Linn (2008), knowledge-engineered rules were developed to score students' experimental behavior for a physics unit on airbags in the learning environment Web-based Inquiry Science Environment (WISE; Slotta & Linn, 2009). The rules were used to analyze three exploration behaviors, namely, total number of trials, trial variability (a measure to determine what range of values for each variable were tested), and experimentation validity (a hand-scored measure of the extent to which CVS was followed and consistent with a chosen investigation question). They found that students who conducted valid experiments (those in which valid inferences could be made from data) tended to conduct fewer trials but learned more content knowledge from the airbags unit as measured by pre/post comparisons. The authors noted that high scores on experimentation validity reflected several skills, including controlling for variables, successfully mapping questions onto experimentation variables, correctly interpreting outcomes, and planning investigations in advance.

These approaches may fail, in our opinion, to properly measure skill in more open-ended environments in which students tend to exhibit a variety of data collection strategies. For example, consider using McElhaney and Linn's (2008, 2010) approach to measure skill at designing controlled experiments by computing successive pairwise CVS trials. This approach may fail to catch "corner cases" in which students exhibit additional behaviors; that is, a student may run repeated trials to observe the microworld, change one variable, run a few more repeated trials, change one variable, and so on. As another example, a student may initially run pairwise experiments and then search for interaction effects. In both cases, students appear to understand how to design controlled experiments but are engaging in other kinds of valid exploration behaviors. The successive pairwise controlled experiments scoring rule, though, would yield a low estimate of skill. The averaged-based approaches of Harrison and Schunn (2004) also would yield lower estimates. As illustrated, because students may collect any data they like and exhibit a variety of strategies, engineering rules and identifying all potential corner cases can be quite difficult.

Our Approach to Inquiry Skill Assessment Using Educational Data Mining

In our detector development work described in this article, we address the skill of *designing controlled experiments* and define it as when a student conducts an experiment that tests the effects of unconfounded comparisons of manipulable variables on outcome variable(s) (Sao Pedro, Baker, Gobert, et al., 2013). This skill is related to the understanding and successful use of the CVS (cf. Chen & Klahr, 1999), a strategy entailing the procedural and conceptual understanding of how, when, and why a controlled experiment should be conducted so that one can make valid inferences about the effects of one independent variable on a dependent variable (Chen & Klahr, 1999; Kuhn, 2005b). The designing controlled experiments skill, as operationalized by us, is different than CVS in that CVS focuses on creating a *single* contrastive and controlled experiment (a single pair of trials) to determine the effects of a variable (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004). Designing controlled experiments, in contrast, applies to the collection of an entire data set and could involve multiple trials and variables. This is reflected in our assessment metric as well, which takes multiple trials into account, unlike other assessment/scoring techniques that score trials as CVS *if and only if* they are collected sequentially (Chen & Klahr, 1999; Klahr & Nigam, 2004; McElhaney & Linn, 2008, 2010). Because students do not conduct trials in a sequential, lockstep fashion (a parameter used in other assessments of the skill), it was important for us to use prior findings about students' strategies/approaches in order to develop the hand-scoring categories upon which our designing controlled experiments detector is based (more on this later in the article).

Here we describe a potential solution to many of the design and measurement issues for the designing and conducting experiments skill, situating our work within our inquiry learning and assessment environment, Inq-ITS (www.inq-its.org). We use techniques from educational data mining (cf. Baker & Yacef, 2009; Romero & Ventura, 2010), which, broadly described, is a set of powerful methods for analyzing patterns in educational data. Several different techniques exist, from more exploratory and bottom-up (i.e., data-driven) techniques to more top-down (i.e., theory-driven) techniques. Educational data mining has been used for a variety of goals: to study which types of interventions are most effective (cf. Beck & Mostow, 2008; Chi, VanLehn, & Litman, 2010), to refine domain knowledge models (Cen, Koedinger, & Junker, 2008; Desmarais, Meshkinfam, & Gagnon, 2006; Pavlik, Cen, & Koedinger, 2009), to build automated detectors of relevant constructs during student learning (Baker, Corbett, Roll, & Koedinger, 2008; Cetintas, Si, Xin, & Hord, 2010), and to engage in formative and performance assessment (Gobert et al., 2012; Mislevy et al., 2012). Educational data mining can be a powerful method; however, pedagogically meaningful findings are most likely yielded when data mining is guided by theoretical principles, in this case about both

inquiry learning and the prior literature on students' inquiry learning. Although educational data mining, particularly in exploratory data mining, appears to be distinct from the forward-design processes used in the psychometric community (Mislevy et al., 2012), we argue that our approach, in which categories are developed a priori to guide data mining, can lead to valid metrics for the development of assessment models.

Our approach is unique in that we use text replay tagging to provide ground truth indicators of scientific inquiry skill that can be used to model these skills. In contrast to knowledge engineering methods, in which rules are defined rationally (cf. Koedinger & MacLaren, 2002), this method enables us to leverage human assessments of scientific inquiry skill and then build models that can replicate those human assessments. Knowledge-engineered models have some risk of oversimplifying ill-defined skills and phenomena, as mentioned earlier (Harrison & Schunn, 2004; McElhaney & Linn, 2008, 2010). Text replays, discussed in further technical detail later, are a method for quickly annotating log files by hand (Baker & de Carvalho, 2008; Baker, Mitrovic, & Mathews, 2010). In this approach, a set of categories of interest is developed from the prior literature and/or cognitive task analysis (Ericsson & Simon, 1980). In our studies, the development of categories for hand-scoring is guided by analyses of students' experimenting in microworlds as either systematic or haphazard (Buckley, Gobert, Horwitz, & O'Dwyer, 2010); the categories are further refined based on the national inquiry frameworks (NRC, 1996) and the literature on students' difficulties conducting inquiry (described earlier).

Next machine learning is used to produce a *detector,* or a model that can identify specific behavior within a student's real-time use of the learning software, for each coding category. After a detector is developed, it is validated for reliability for new students and new curricular materials and can then be used to auto-score students' data. Later we provide more details on our methods and give evidence of their reliability and validity. This approach has been previously used to assess whether a student is gaming the system (Baker & de Carvalho, 2008; Baker, Mitrovic, et al., 2010) and to assess novice programmers' confusion (Lee, Rodrigo, Baker, Sugay, & Coronel, 2011). In brief, educational data mining provides a set of techniques, when appropriately guided, for analyzing process data (Rupp et al., 2010) about students' learning such that rich description can result.

METHOD

Participants

Participants were 148 eighth-grade students, ranging in age from 12–14 years, from a public middle school in central Massachusetts. Students belonged to one

of six class sections and had one of two science teachers. They had no previous experience using microworlds within our learning environment, Inq-ITS (www.inq-its.org).

Materials

Inq-ITS (www.inq-its.org; formerly www.scienceassistments.org) is a Web-based intelligent tutoring and assessment system for physics, life science, and earth science that aims to automatically assess (and in the future, automatically scaffold) scientific inquiry skills in real time within interactive microworld simulations. These simulations, designed for use at the middle school level, span several science domains, including physical, life, and earth science (Gobert et al., 2012; Gobert, Sao Pedro, Toto, Montalvo, & Baker, 2011). Each microworld targets domain-specific concepts defined in the Massachusetts curricular frameworks content standards for middle school science (Massachusetts Department of Education, 2006). Within each microworld, inquiry skills identified in the *National Science Education Standards* (NRC, 1996) for middle school are assessed, including hypothesizing, designing and conducting experiments, interpreting data, warranting claims, and communicating findings. Because the focus of the present work is on assessing data collection skills within two of our physical science microworlds, the phase change microworld and the density microworld, we describe these microworlds in more detail here.

Phase Change Microworld Activities. The phase change microworld, shown in Figures 1 and 2, aims to foster understanding about the melting and boiling processes of a substance through semistructured scientific inquiry activities. A typical activity provides students with an explicit goal to determine whether one of four variables (container size, heat level, substance amount, and container covered) affects properties of a substance's phase change (melting point, boiling point, time to melt, and time to boil). Students address the goal by hypothesizing, collecting data, reasoning with tables and graphs, analyzing data, and communicating findings about how a variable affected the outcomes.

Structuring each of the previously mentioned inquiry tasks into different phases (i.e., observe, hypothesize, experiment, and analyze data) supports students' overall experimentation. Students begin in the hypothesize phase and are allowed some flexibility to navigate between phases. In the hypothesize phase, students use the hypothesis construction tool (see Figure 2) to generate testable hypotheses. The hypothesis tool is a fill-in template that enables students to construct hypotheses, such as "If I change the container size so that it increases, the melting point stays the same." If students are not yet ready to specify a hypothesis, they can first explore how the simulation works by navigating to the observe phase. In the observe phase, students are simply presented the simulation and can design and

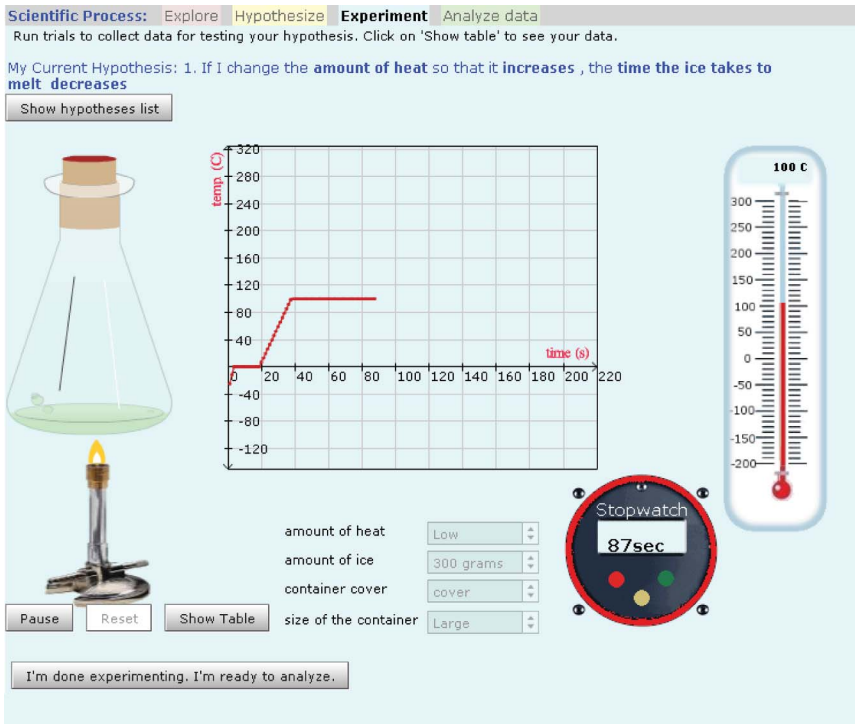

FIGURE 1 Phase change simulation. In this phase of inquiry, students collect data to test their stated hypotheses (color figure available online).

run as many experiments as they like, observing how the simulation changes over time.

Once students specify one or more hypotheses, they can choose to begin collecting data to test those hypotheses by navigating to the experiment phase. The experiment phase (see Figure 1) is similar to the observe phase in that the students design and run experimental trials. However, in the experiment phase, two inquiry support tools—a data table summarizing previously run trials and a hypothesis list—are provided to help students plan which experiments to run next and to help them monitor their progress. Finally, once students feel that they have collected sufficient data to test their hypotheses, they can navigate to the analyze phase. In this phase, students are shown their data and use the data analysis construction tool (similar to the hypothesis tool) to construct an argument that supports or refutes their hypotheses based on the data they gathered.

Because the environment provides a moderate degree of student control, students have some freedom to navigate between inquiry phases and have flexibility

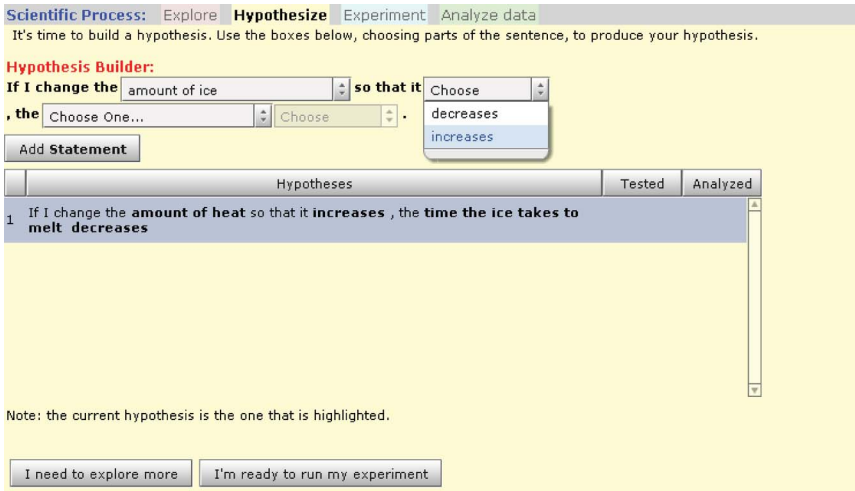

FIGURE 2 Hypothesizing tool for the phase change microworld (color figure available online).

within each phase to conduct many actions. For example, while in the hypothesize phase (see Figure 2), students can elect to explore the simulation more before formulating any hypotheses by moving to the observe phase. Alternatively, they can choose to specify one or more hypotheses before collecting data. Within the experiment phase (see Figure 1), students can run as many experiments as they wish to collect data for any one or all of their hypotheses. Within the analyze phase students also have several options. As they construct their claims, students can decide to go back and collect more data, or, after constructing claims based on their data, they can decide to create additional hypotheses, thus starting a new inquiry loop.

Density Microworld Activities. The density microworld, shown in Figure 3, enables students to inquire about the relationships between mass, volume, and density and is based on Archimedes's principle of buoyancy. Similar to phase change, a typical task in this domain provides students with an explicit goal to determine whether a particular independent variable (orientation of object, type of liquid, volume of object, and mass of object) affects density. However, unlike phase change, the activities are more open ended in that they have fewer inquiry support tools.

More specifically, there are three differences between the phase change and density activities. First, support tools for hypothesizing and analyzing data are not provided. Students instead write hypotheses and analyses in open response boxes. Second, the manner in which students move between phases of inquiry is

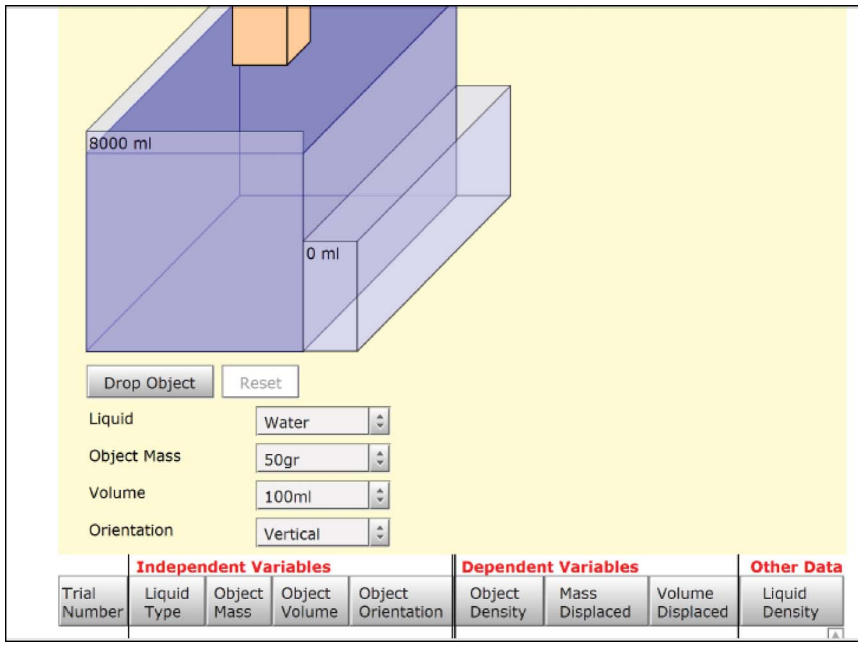

FIGURE 3 Density microworld—Archimedes (color figure available online).

slightly different. Unlike in phase change, students engage in hypothesizing only once. Third, students cannot elect to observe before forming a hypothesis. These differences exist because the inquiry support tools and navigation components were not implemented for density at the time we collected our data. Though this was the case, this design enabled us to examine whether and how students' inquiry differed without the inquiry widgets (we are currently analyzing these data).

To provide a more concrete example, one density activity gives students the goal of determining whether the orientation of an object (whether it is dropped into water horizontally or vertically) will affect its density. The students first write their hypothesis pertaining to that variable. Next they are presented with the simulation in which they design and run experiments to collect data. Finally, after signaling that they have collected sufficient data, students write their analyses. However, when writing their analyses, students can still use the simulation to collect more data as needed if they believe such data would aid them in drawing conclusions.

Procedure

All students engaged in phase change and density learning activities. However, the order in which students did these two activities was randomly assigned by the

Science Assistments system. Therefore, some students received density activities first, whereas others received phase change activities first. This randomization occurred at the student level; students within the same class conducted inquiry in each domain (density first or state change first) in different orders. This design reduced possible effects due to classroom culture, teacher, and so on.

For each domain, students engaged in a series of activities aimed at assessing content knowledge and inquiry skill. These activities progressed as follows:

Domain Content Knowledge Pretest. This test consisted of multiple-choice and fill-in questions that measured incoming knowledge about the domain. Because the focus of this work is on assessing behavior within microworld-based activities, we do not discuss the pretests in further detail.

Microworld-Based Inquiry Activities. Each activity was designed such that students were provided with a goal orientation (described in the task description) for each inquiry phase. Specifically, the task description asked students to hypothesize, design and conduct experiments to test their hypotheses, interpret their data, warrant claims, and communicate findings. As students engaged in inquiry, the Science Assistments system logged all interactions with the microworlds and activities. Each domain's activities are described in more detail as follows.

Phase change activities. One exploratory activity and four inquiry activities were administered using the phase change microworld with inquiry support tools. In the exploratory activity, students familiarized themselves with the phases of matter microworld and concepts. Then students progressed to the four inquiry activities. Each activity asked students to investigate how a particular independent variable (e.g., container size) affected all of the dependent measures (e.g., melting point, time to boil). Students engaged in inquiry as described in "Materials" to address each goal as directed by orienting tasks. A different goal was given in each of the four activities, one for each independent variable. However, students could choose to ignore the current goal and test any hypothesis and conduct as many different kinds of experiments as desired.

Density activities. One exploratory activity and three inquiry activities were given in the density microworld. Similar to phase change, the exploratory activity enabled students to familiarize themselves with the microworld. The three activities that followed asked students to investigate how a particular independent variable (e.g., the orientation of an object) affected density. Students then engaged in inquiry as described in "Materials" to address each goal. Again, students could choose to ignore the given goal.

Domain Content Knowledge Immediate Posttest. After completing the microworld-based activities, students answered identical content posttest items.

BUILDING AND VALIDATING A MACHINE-LEARNED DETECTOR FOR THE DESIGNING CONTROLLED EXPERIMENTS BEHAVIOR

Our goal is to develop and validate machine-learned detectors of the designing controlled experiments behavior that can correctly predict behavior within microworlds for two domains, phase change and density. Previously we developed a detector of this skill that could predict whether a student was designing controlled experiments, and we found that it could successfully predict this skill within the phase change microworld (Sao Pedro, Baker, Gobert, et al., 2013). The detector was built using text replay tagging, as mentioned previously. In this approach, human coders label segments of students' log files with behaviors (e.g., designing controlled experiments). These codes are checked for interrater reliability. Then the judgments (also called *labels*) are combined with features (also called *attributes of the data* or *predictor variables*) of the student interaction within the clip and are given to a machine learning algorithm to discover a model—that is, to identify which features can predict the human judgments.

Though the detector worked well (e.g., successfully predicting text replays coded for entirely new students), we noticed a drawback in the resultant model. When inspecting the detector more closely, we noticed that it did not contain features considered theoretically important to the behavior (Buckley et al., 2006; Chen & Klahr, 1999; McElhaney & Linn, 2010). For example, a feature that tracked the number of controlled comparisons a student made in his or her data set was not present in the final model (Sao Pedro, Baker, Gobert, et al., 2013), though this is a known indicator of skill at designing controlled experiments (Chen & Klahr, 1999; McElhaney & Linn, 2010). In addition, other features that were *not* considered important were present in the model, for example, a feature related to whether the student chose to hide (or display) the hypothesis list during inquiry (Sao Pedro, Baker, Gobert, et al., 2013). We believe that, despite predicting the labels, the detector did not fully reflect the designing controlled experiments skill, meaning that it had low construct validity. In turn, we believe this may negatively have impacted its predictive performance. Since then, we have reengineered the designing controlled experiments behavior detector to increase its construct validity using the same data set in the context of the phase change activities. This redesign required us to follow a slightly different procedure than our original text replay tagging approach, particularly in how we built and validated the detector. Thus, we present our modified approach here. Then we determine whether this detector, built for phase change, can be used as is to correctly classify behavior within the other domain, density.

Overview of the Text Replay Tagging Process

As shown in Figure 4, there are several steps in building and validating the behavior detector for phase change using text replay tagging. Briefly, the process begins by collecting data from students by having them engage in inquiry within the microworld activities. The activities are designed to enable the student to engage in meaningful and appropriate (or inappropriate/ineffective) science inquiry skills/behaviors in understanding the relationship between dependent and independent variables. From there, log files are segmented into meaningful sets of student actions (called *clips*) that help in identifying the behavior of interest, designing controlled experiments. Human coders then tag a subset of these clips with the behavior. A set of features that summarizes the clips is distilled and combined with the labels. The combination of tags and features provides a basis for discovering a model for the designing controlled experiments behavior and testing how well the model performs.

Collecting Student Data and Generating Log Files

Our process for developing data-mined models of student inquiry behavior relies on developing models that can replicate human judgments. Thus, the first step in our process is to collect student data by having students engage in inquiry within our microworld activities and logging all of their interactions. In our original work (Montalvo, Baker, Sao Pedro, Nakama, & Gobert, 2010; Sao Pedro, Baker, et al.,

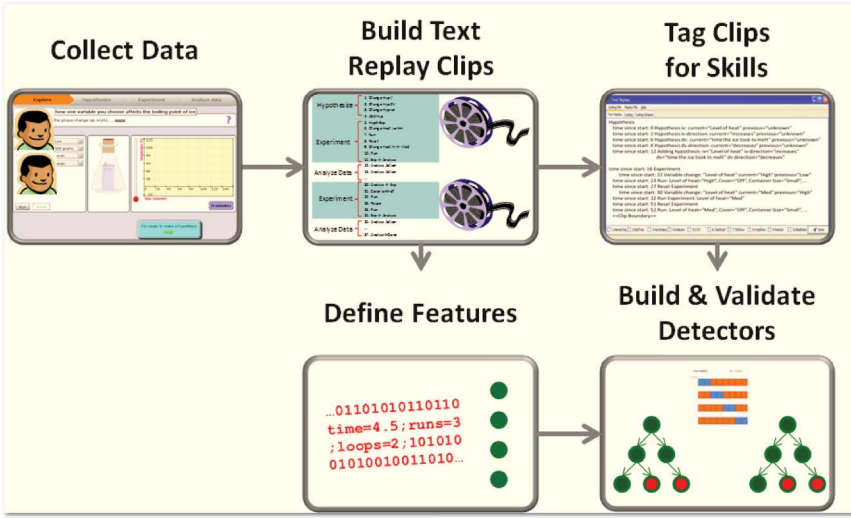

FIGURE 4 Tag clips for inquiry behaviors (color figure available online).

2010; Sao Pedro, Baker, Gobert, et al., 2013), we began by observing students' inquiry within the phase change activities.

Next all log files were distilled. These log files encompassed the entire sequence of steps taken by a student in conducting his or her investigations. The log files contained each student's time-stamped low-level interactions within the simulation and inquiry support tools. Some low-level actions included constructing hypotheses and analyses, changing simulation variables, running experiments, and transitioning between inquiry phases (e.g., moving from hypothesizing to experimenting). A full description can be seen in Sao Pedro, Baker, Gobert, et al. (2013).

Segmenting Log Files Into Clips for Hand-Scoring

Log files for each student and activity are further segmented into contiguous sequences of actions called *clips* (see Figure 5). Clips contain all of the actions necessary to enable a human coder to identify the behaviors of interest (in this case one type of inquiry, designing controlled experiments). Student behavior/skill demonstration is labeled by the coders with reference to clips (i.e., one code per human coder per clip). Our detectors make predictions about student behavior at the same grain size (i.e., one detector prediction is made per clip). In the text replay methodology, an important decision is choosing which actions should be included in a clip. This decision is important because a human coder needs sufficient information to properly identify behavior but should not be burdened with extraneous information that could hinder his or her identification speed and accuracy. In previous work using this methodology, Baker and de Carvalho (2008) and Baker, Mitrovic, et al. (2010) segmented clips according to a prespecified length of time (e.g., all actions that occurred in a 20-s interval). For our purposes, however, we

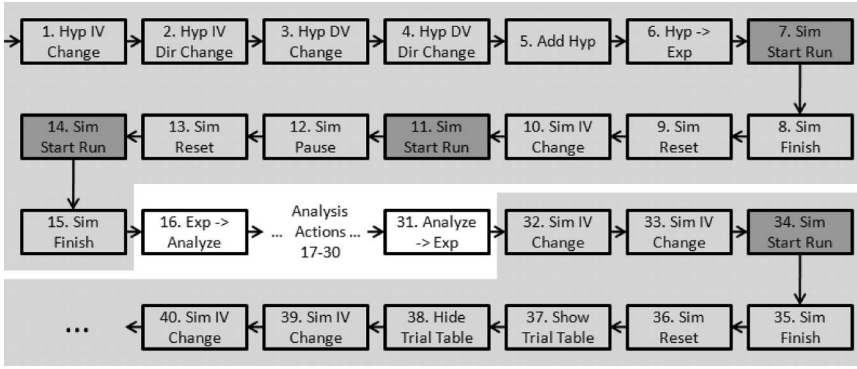

FIGURE 5 Example sequence of student actions for a phase change activity. Two clips (shown in light grey) are generated because the experiment stage was entered twice.

needed to show significant periods of experimentation so that human coders could precisely evaluate data collection behavior. Thus, clips were defined as including all actions from the hypothesize phase (to provide information about which hypotheses were stated) and all actions from the experiment phase (to provide information about which trials the student conducted). In addition, we note that several clips could be generated for one student in a single activity. This occurs when students engage in several inquiry cycles within an activity. Therefore, if a human coder is coding a student's fifth clip, the student's actions for the first through fourth clips are also displayed. This additional information helps human coders disambiguate behavior when a student has been using the microworld for a considerable amount of time (e.g., data collected in a previous inquiry cycle can be utilized by students to test a hypothesis in a later inquiry cycle). Similarly, a student may run a single trial in a clip that, when combined with trials from prior clips, indicates skillful inquiry. When one looks at such a clip in isolation, a single trial by itself would appear like haphazard inquiry (Kuhn et al., 1992), and thus experimentation skill could be misjudged. Thus, without showing the full history related to hypothesizing and experimenting, human coders would not be able to recognize students' sophisticated inquiry behaviors, in turn possibly leading to incorrect tagging and assessment. In summary, a clip contains all student actions relevant to hypothesizing and experimenting and takes into account all actions from earlier clips within the same activity.

Tagging Behavior in Clips for Phase Change Activities

The next step is for coders to label clips with behaviors of interest. In our previous work (Sao Pedro, Baker, Gobert, et al., 2013), we identified 10 labels. In this process, human coders apply behavior tags to a text-based representation of key information in a clip, called a *text replay.* A text replay, shown in Figure 5, summarizes clip actions and highlights important aspects of students' inquiry processes. For example, students' full experimental designs are displayed. In addition, fully stated hypotheses are also shown.

To illustrate how behavior is labeled by a human coder, consider the text replay shown in Figure 6, which was tagged in part as demonstrating the designing controlled experiments skill. To tag this, the human coder focused primarily on the trials run by the student. In this experimentation cycle, the student ran a total of three trials, as indicated by the "microworld run" statements at time 94 s, 117 s, and 140 s. For each trial, the student changed only the level of heat variable in a successive manner, comparing a low level to medium, and then medium to high. The student spent 55 s doing so. In the second experimentation cycle, the student did not collect any additional data. Because of the consistency in manipulating only one variable at a time between trials, we would label this clip as demonstrating the designing controlled experiments.

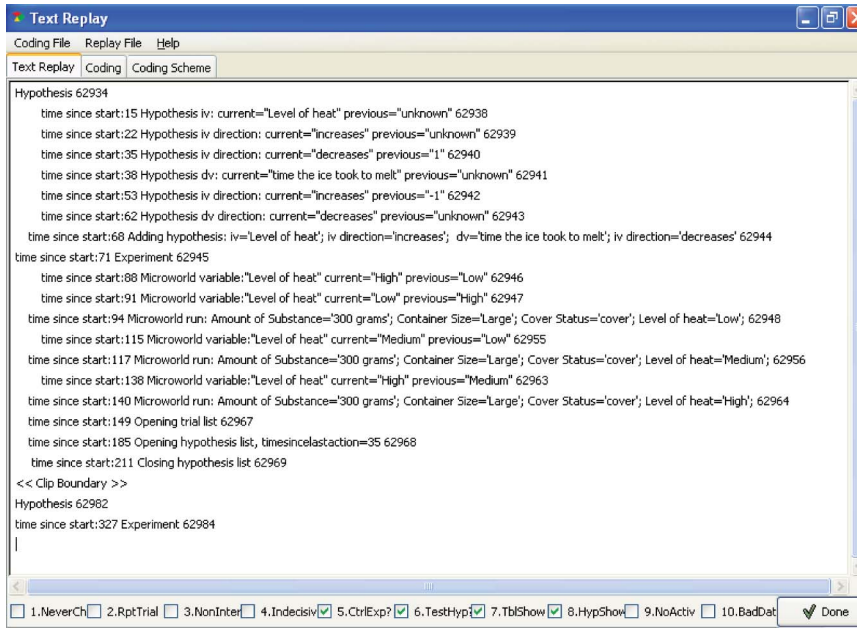

FIGURE 6 Text replay tagging tool for the phase change environment. Each line indicates a student action, with the time when it occurred and details of the student action. This clip was tagged as involving designing controlled experiments. This was identified by the coder, seeing that the student ran three trials (at 94, 117, and 140 s, respectively, after the start of the clip), only changed one variable between the first and second trials (level of heat from low to medium), and only changed one variable between the second and third trials (level of heat from medium to high) (color figure available online).

At this point, we have described a process for segmenting students' low-level actions into clips and labeling clips with behaviors. The machine learning approach we use necessitates the use of *clips tagged with behaviors* and a *feature set,* which summarizes clips. A machine learning algorithm attempts to find which combination of features and feature values best predicts the behaviors given the data.

Data Sets for Building and Validating a Detector Within Phase Change Activities

Human coders tagged many clips to train, refine, and test our detector for the designing controlled experiments behavior. Typically, one data set (e.g., one set of clips) is divided into subsets, and then each subset is repeatedly used to train machine-learned detectors. Each subset is omitted once from the training set and

instead is used to test the detectors in a process termed *cross-validation* (cf. Efron & Gong, 1983). This approach was used in our prior work building behavior detectors (see Sao Pedro, Baker, et al., 2010; Sao Pedro, Baker, & Gobert, 2013a). However, in this work, we used an alternative method, which yielded a detector with superior construct validity. This method required different data sets to be used to train, refine, and test the detector. These data sets were as follows.

Training Set (601 Clips). In our initial work (Sao Pedro, Baker, Gobert, et al., 2013; Sao Pedro, Baker, et al., 2010), two human coders had tagged 571 clips for training the behavior detector. Because several clips could be generated per activity, a single, randomly chosen clip was tagged per student per activity. This ensured that all students and activities were equally represented in this data set. To validate that a replicable construct was being coded, interrater reliability was established across two human coders. Prior to testing for interrater reliability, the two coders discussed the coding scheme and coded several clips together. Next, to test for interrater reliability, both coders tagged the same 50 clips. Afterward, the remaining clips were equally divided for each coder to code independently. It took roughly 1 hr for a human coder to tag 60 clips with behaviors (Sao Pedro, Baker, Gobert, et al., 2013). Interrater reliability for the tags was high overall: The kappa for designing controlled experiments was 0.69, 69% better than chance, a level of agreement sufficient in the past to successfully develop text replay–based behavior detectors (Baker & de Carvalho, 2008; Baker, Mitrovic, et al., 2010; Lee et al., 2011).

By chance, the random-selection sampling method produced few clips representing students' first data collection within an activity. Building a detector from this imbalanced data set could have impacted its performance because it may not have properly learned how to classify such clips from so few training examples. To have a more representative training set, one human coder tagged an additional 30 randomly selected "first" clips. In total, 31.4% of the clips were tagged as designing controlled experiments.

Validation Set (100 Clips). A special set of clips was tagged by one of the human coders who tagged for the training set. This set was used for engineering the new detector with improved construct validity (described in more detail later). This set contained 20 randomly chosen first clips, 20 randomly chosen second clips, and so on, up through 20 randomly chosen fifth clips. Unlike in our prior work (Sao Pedro, Baker, Gobert, et al., 2013), we did not stratify clips by student or activity when selecting clips for this set. Stratification, an accepted practice in testing educational data mining models (cf. Baker et al., 2008; Pardos, Baker, Gowda, & Heffernan, 2011; Sao Pedro, Baker, & Gobert, 2013b), attempts to ensure that students (or students' activities) are equally represented so that the goodness estimates of data-mined models/detector are not biased. We could not

perform this stratification because all students and activities were used to build the training set; thus, stratification would not have removed biases already present in this data set. In total, 34.0% of the clips were tagged as designing controlled experiments.

Held-Out Test Set (439 Clips). The same human coder who tagged clips for the validation set also tagged an additional 439 clips for the held-out test set (i.e., clips not used to build or refine the machine-learned model but instead used solely to test the model's goodness; cf. Witten & Frank, 2005). These clips were used to test how well the detector could predict behavior for unseen data in the phase change activities. A total of 64.7% of the clips in this data set were tagged as designing controlled experiments.

Feature Set Used to Summarize Clips

We distilled a set of features (also known as *attributes* or *predictor variables*) to summarize the actions that occurred in the clips. These features were derived from prior work on building other behavior detectors (e.g., Baker & de Carvalho, 2008; Baker, Mitrovic, et al., 2010; Walonoski & Heffernan, 2006), prior work on building cognitive models of experimentation behavior (Buckley et al., 2006, 2010; McElhaney & Linn, 2010; Schunn & Anderson, 1999), and prior cognitive science research on experimentation (Kuhn et al., 1992; van Joolingen & de Jong, 1991, 1993) and the CVS (cf. Chen & Klahr, 1999). Our original designing controlled experiments detector considered 73 features (Sao Pedro, Baker, Gobert, et al., 2013). Since then, we refined our feature set and selected 11 candidate features for potential inclusion into the model. These features were identified using a combination of predictive power and construct validity. Construct validity was used to winnow out features without a theoretical justification for connection to designing controlled experiments. Predictive power was used by exploring which features had a correlation of 0.2 or higher with the designing controlled experiments behavior. This exploration was performed using the training set data only. By having a held-out test set, we could reduce the potential of overfitting stemming from this feature selection process. These features, computed per clip, were as follows:

- *All actions count:* A count of all low-level actions found in a clip. Recall that these are actions in the hypothesize and experiment phases of inquiry. These actions include changing variables when making hypotheses; adding hypotheses; running, pausing or resetting the simulation; changing simulation variables when designing experiments; and finally displaying or hiding the data table and hypothesis list (Sao Pedro, Baker, Gobert, et al., 2013).
- *Complete trials count:* The number of trials in which the student ran the simulation to completion (i.e., without restarting the trial).

- *Total trials count:* The total number of trials started within the clip, regardless of whether the student allowed the simulation to run to completion.
- *Simulation pause count:* The number of times the simulation was paused.
- *Simulation variable changes count:* The number of times the values of simulation variables were changed (i.e., the student toggled the simulation variable dropdown boxes shown in Figure 2) while the student was designing experiments.
- *Simulation variable changes count related to stated hypotheses:* The number of times the values of simulation variables explicitly stated in hypotheses were changed.
- *Number of pairwise repeated trials:* A count of the pairs of trials that had identical experimental setups. This count considers *any* two trials in the entire clip.
- *Number of successive repeated trials:* The same as the pairwise count, except that only adjacent (successive) trials (e.g., between Trials 2 and 3, between Trials 4 and 5) are considered.
- *Number of pairwise controlled trials, with repeats:* A count of the pairs of trials in which exactly one simulation variable (independent variable) had different values between trials, and all other variable values were identical (cf. Chen & Klahr, 1999). Because it is a pairwise count, any pair of trials is considered. Furthermore, if any trial is a repeat of an earlier trial, it is still considered in this count.
- *Number of successive controlled trials, with repeats:* Same as the pairwise controlled trial count, except that this count only considers successive trials.
- *Number of pairwise controlled trials, ignoring repeats:* Same as the pairwise controlled count previously mentioned, except that if a trial is a repeat of an earlier trial, it is not considered.
- *Number of successive controlled trials, ignoring repeats:* Same as the previous feature, except between two successive trials.

We note that the counts for constructing controlled trials, constructing repeat trials, and changing variables associated with stated hypotheses are all features used by others to directly measure procedural understanding associated with designing controlled experiments (Buckley et al., 2006, 2010; Chen & Klahr, 1999; McElhaney & Linn, 2010). The other features, though not directly related, may also help distinguish procedural understanding. Thus, they are also candidates we considered when building our machine-learned behavior detectors.

For each clip in the three data sets (training, validation, and held-out test), the ground truth labels produced by the human coders were connected to the corresponding feature values. With these data, we could now build and validate the machine-learned detector.

Building the Machine-Learned Detector for Designing Controlled Experiments

We utilized the training and validation data sets to build and refine our designing controlled experiments detector. The output of this process is a detector that takes as input a clip (not a student) and, by looking at feature values of the clip, yields a prediction of whether the designing controlled experiments behavior was demonstrated in the clip. Our detector was developed using RapidMiner 4.6 (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006).

A candidate machine-learned detector was built using the entire training set and a subset of the features. These data were inputted to a J48 decision tree algorithm, an open-source implementation of the C4.5 decision tree algorithm (Quinlan, 1993). Decision trees are a type of prediction model that repeatedly divide data based on the values of a variable, creating branches that can look different from one another. For instance, a tree may branch on the amount of time taken by the student per action. If the student averages less than 10 s, then a left branch is considered, whereas if the student averages more than 10 s, a right branch is considered. On the left branch, the model may again branch, considering whether the student repeats the same response. If the student does so, the model goes down the left-left branch; if not, the model goes down the left-right branch. In contrast, on the right branch, a different variable may govern the choice between the right-right branch and the right-left branch. Different decision tree algorithms use different procedures for selecting which variables to branch on and when to stop branching. When branching stops, a leaf node is reached that has a final model prediction and typically a confidence for that prediction.

J48 decision trees make predictions on categorical or binary data (and thus J48 is referred to as a *classification algorithm*). J48 trees provide confidences (probabilities) for their predictions, as well as overall predictions. Branch variables are selected using the information gain metric (cf. Kent, 1983), which indicates the degree of uncertainty reduction achieved by selecting that variable. J48 decision trees are thought to be particularly good at reducing overfitting (e.g., fitting noise rather than signal) by using a post hoc pruning approach that reduces tree complexity (Quinlan, 1993). At the same time, they can capture complex nonlinear relationships in the data that are difficult to see with logistic regression, for instance. (Linear regression is not used because a binary variable is being predicted, which is a violation of the assumptions of linear regression.) The J48 algorithm has two parameters that can be set to control for overfitting to the training data. We set the minimum number of data points per leaf (M) to 2 and the confidence threshold for pruning (C) to 0.25. This algorithm was chosen because of its previous success in building behavior detectors in educational software (Baker & de Carvalho, 2008; Walonoski & Heffernan, 2006) and because the decision trees generated are relatively interpretable by humans.

To home in on an optimal machine-learned detector built from the best features, we performed a manual backward elimination search (cf. Witten & Frank, 2005) as follows. We first built a model considering all 11 features. A domain expert ordered the set of remaining features in terms of the theoretical support for them. Then features were removed one at a time, starting with the feature with the least theoretical support. At a given point in the search, the domain expert also explored several alternatives by eliminating several features, still one at a time. For each candidate feature set, a decision tree was constructed using the training set (yielding a candidate model). Then we tested the model's predictive performance on the validation set of 100 clips.

Predictive performance was measured using A' (Hanley & McNeil, 1982) and kappa. Briefly, A' is the probability that when given two clips, one labeled as demonstrating a behavior and one not, a detector will correctly identify which clip is which. A model with an A' of 0.5 performs at chance, whereas 1.0 indicates that it performs perfectly. Cohen's kappa assesses whether the detector is better than chance at labeling behavior. A kappa of 0.0 indicates chance-level performance, whereas 1.0 indicates perfect performance. These metrics were chosen because they attempt to compensate for successful classification occurring by chance (Ben-David, 2008). Both metrics are used because they have different tradeoffs. A' can be more sensitive to uncertainty but looks at the classifier's degree of confidence; kappa looks only at the final label, leading to more stringent evaluation. In addition, we consider two other metrics in the final model evaluation: precision and recall (cf. Witten & Frank, 2005). *Precision* is the number of clips correctly labeled as a particular behavior, such as designing controlled experiments (true positives), divided by the total number of clips labeled as positive by the detector (the sum of true positives and false positives). False positives are items that are incorrectly labeled as belonging to a particular behavior. *Recall* is defined as the number of true positives divided by the total number that actually belong to the class (the sum of true positives and false negatives). Here, false negatives are items that were not labeled as belonging to the particular class but should have been. It is worth briefly mentioning two metrics that are not used in this article, and why. First, correlation is not used, as we are predicting binary labels rather than numerical variables. Using correlation with binary variables is generally seen as a violation of the mathematical assumptions of this metric. Second, overall accuracy (also termed *percent correct*) is not used. This is because accuracy does not take the actual frequency of categories in the data into account; for instance, if students design controlled experiments 90% of the time, then a model that always indicates that a student is designing controlled experiments will achieve an accuracy of 90%. Kappa is increasingly used as an alternative to accuracy, which takes base rates into account.

To home in on an optimal model, we repeated the above process, removing one feature at a time, until performance degraded. In other words, we continued to

remove features as A' and kappa increased but stop when they decreased. In some cases, A' and kappa may not have increased or decreased together, or they may not have changed at all. When we compared two candidate models, the model with the higher kappa was preferred. However, if A' decreased greatly and kappa increased slightly, the model with higher A' was chosen. If two models yielded the same values, the model with fewer features was chosen.

The best performing decision tree for the designing controlled experiments detector utilized eight features. These features were complete run count, variable changes made when designing experiments, changes to variables associated with stated hypotheses when designing experiments, adjacent and pairwise controlled experiments counts (both with and without considering repeats), and pairwise and adjacent repeat trials counts. This new decision tree was more compact and more easily interpretable than the original model presented in Sao Pedro, Baker, Gobert, et al. (2013). The decision tree had a total of 19 nodes and 10 leaves. The root of the tree (the first decision made in determining whether a student's experimentation reflected skill at designing controlled experiments) was a terse measurement of the construct, the number of adjacent controlled trials (repeat trials included). If students never ran two controlled experiments in a row, the model predicted that the students did not know how to design controlled experiments with 99% confidence. From there, the model yielded predictions by integrating pairwise controlled trial counts, number of simulation variable changes, and number of times the student ran the simulation to completion. For example, one such rule included the following:

IF count of adjacent controlled experiments (with repeats) $= 1$ AND

count of simulation variable changes ≤ 2 AND

count of pairwise controlled experiments (with repeats) > 1 AND

complete trials count > 2

THEN predict that the clip is a demonstration of designing controlled experiments with 73% confidence.

Intuitively speaking, this extra complexity learned from student data attempts to disambiguate situations in which students ran many or few trials, those corner cases that are typically hard to codify by hand.

Against the validation set, the detector had A' $= 1.0$ and $\kappa = .84$. We note that these values are extremely high and that this should be expected because we maximized prediction against the validation set during our search for the best model. Because we benchmarked goodness against the validation set so many times, it is likely we overfit our models to the validation set, meaning that they would not predict well for unseen student data. Thus, we performed an additional validation

set of testing the best model against data not used in constructing detectors or searching for a best model. We describe the results of this additional validation step in the next section.

Validating the Behavior Detector Within the Phase Change Environment

In validating the detector for phase change, we aim to measure how well it can predict the designing controlled experiments behavior in unseen clips, clips that were not used to build or select the best candidate model. Thus, we validate it by determining how well it could predict behavior for clips in the held-out test set. As mentioned earlier, typically validation is performed using cross-validation over all labeled data (cf. Efron & Gong, 1983). However, we used the entire training set to determine an initial set of features and used the validation set to reduce that feature set. Therefore, the held-out test set was used to reduce bias in the estimation of the detector's goodness.

The matrix showing raw agreement between the detector's prediction of whether actions in a clip demonstrated designing controlled experiments and the human coder's tagging is shown in Table 1. Overall, we found that the detector was very good at distinguishing a clip that demonstrated the behavior from one that did not, indicated by A' = .94 (i.e., the model could distinguish when a student was designing controlled experiments 94% of the time). In addition, the detector was good at predicting the correct label for a clip as measured by κ = .45. We do note, however, that this detector appears to bias toward inferring that a student is not demonstrating this skill. This is indicated by a lower recall value (.58) than precision (.95). Overall, the performance of this detector is comparable to that of detectors of gaming the system (e.g., Baker & de Carvalho, 2008;

TABLE 1
Classifier Matrix for the Designing Controlled Experiments Detector on the Held-Out Test Set

	Designing	*Controlled Experiments*
	True N	True Y
Pred N	146	118
Pred Y	9	166
	Precision = .95	Recall = .58
	A' = .94	κ = .45

Note. Pred N = The detector claimed that the clip did not reflect skill at designing a controlled experiment. True N = The human coder labeled a student's clip as not designing a controlled experiment. Pred Y = The detector claimed that a clip reflected skill at designing a controlled experiment. True Y = The human coder labeled a student's clip as designing a controlled experiment.

Baker, Mitrovic, et al., 2010). Detectors substantially less accurate than this have served as the basis for intervention that significantly improved targeted students' learning (Baker, Corbett, & Wagner, 2006). This detector achieves comparable or better A' than models considered to be sufficient for use in medical diagnosis for life-threatening conditions such as heart disease and Parkinson's disease (two such examples are given in Fan, Upadhye, & Worster, 2006; and Yanamandra et al., 2011).

As an additional analysis, we confirmed that our new detector with improved construct validity outperformed our original detector described in Sao Pedro, Baker, Gobert, et al. (2013). We did so by testing the predictive performance of the original detector on the held-out test set. We found that the new detector with improved construct validity outperformed the original detector on all metrics: new detector, A' = .94, κ = .45, precision = .95, recall = .58; original detector, A' = .89, κ = .30, precision = .90, recall = .46. Thus, the new detector was better at classifying the behavior, also implying that improving construct validity also improved predictive performance.

Applying Our Detector to Density Data

We now determine whether the designing controlled experiments behavior detector developed for phase change can successfully assess this skill in our density activities. Testing the transferability of the machine-learned detector provides another level of validation that the model does indeed capture the behavior. This could also increase confidence that the detector can identify the behavior in all of our physical science microworlds, which all have similar designs to the two studied here. Note that this analysis does not imply that student scientific inquiry skill is fully domain general; a student may be successful at inquiry skill in one domain and not in another. However, this analysis investigates whether the same model can be used to assess student skill in these domains. To achieve this goal, we again used the text replay tagging methodology for the density data.

Generating Density Clips

Similar to phase change, we distilled log files from students' interactions within three density activities and segmented them into clips. The clips, however, differed from the phase change clips in two ways. This was because the density activities did not utilize the inquiry support tools or inquiry structure. One difference was that students engaged in one single experimentation loop per activity. Therefore, exactly one clip was generated per activity. Also, students did not use the hypothesizing widget to specify their hypotheses; they instead wrote them as open text responses. Thus, a student's process in generating hypotheses was not included.

Tagging Behavior in Clips for Density Activities

Text replays were generated from the density clips so that human coders could tag them with the designing controlled experiments behavior. The format of the text replay was nearly identical to that of the phase change text replays. The main difference was that we included students' written hypotheses to assist the human coders in labeling behavior.

Hand-Scored Density Test Set

One human coder tagged 213 clips generated from 71 students' interactions with the density activities. These students engaged in density activities before seeing any phase change activities. This subset of students was chosen because these clips had been previously tagged for use in other research (Gobert et al., 2012). These students completed three density activities. To ensure that the behavior was properly identified in this new domain, a second coder who tagged clips for phase change also tagged 50 clips for density to test for agreement. Interrater reliability was comparable to that in the previous work in phase change ($\kappa = .66$). In total, 55% of the clips were tagged as demonstrating the skill of designing controlled experiments. As done previously, the same features were computed for each density activity to summarize clips. These were combined with the behavior tags to form the density test set. This corpus of clips was used to determine how well the phase change behavior detector could correctly label behavior in this domain.

Validating the Detector for the Second Domain

To analyze the generalizability of the detector to this second domain, we aim to measure how well it can predict the designing controlled experiments behavior in the density activities. In other words, we revalidate the detector by determining how well its predictions match the human-coded tags for clips in the density test set. As with the phase change microworld, we assess the detector using A' (Hanley & McNeil, 1982), kappa, precision, and recall.

The classification matrix shown in Table 2 shows that agreement was high overall between the predictions and human-coded tags as indicated by the agreement measures. The detector was successful at distinguishing a clip demonstrating the behavior of designing controlled experiments from a clip in which the student did not demonstrate that behavior (A' = .82), signifying that the detector could make this distinction 82% of the time. Also, the detector was quite good at predicting the correct label for a clip ($\kappa = .55$), indicating that the detector was 55% better than chance in terms of agreement with the human coder. As in the phase change data, we note that the detector appears to bias toward inferring that a student is

TABLE 2
Classification Matrix for the Designing Controlled Experiments Detector and Hand-Scoring for Density

	Designing	*Controlled Experiments*
	True N	True Y
Pred N	89	41
Pred Y	8	75
	Precision = .90	Recall = .65
	A' = .82	κ = .55

Note. Pred N = The detector claimed that the clip did not reflect skill at designing a controlled experiment. True N = The human coder labeled a student's clip as not designing a controlled experiment. Pred Y = The detector claimed that a clip reflected skill at designing a controlled experiment. True Y = The human coder labeled a student's clip as designing a controlled experiment.

not demonstrating skill. This is indicated by a lower recall value (.65) than precision value (.90). As shown by the high A' and kappa values, the detector can successfully predict whether a clip demonstrates the designing controlled experiments behavior in the density activities. These findings provide evidence of the detector's broader applicability to assess this skill in another physical science domain.

SYNTHESIS, IMPLICATIONS, AND FUTURE WORK

Rationale and Goals for the Article

In the past, students' log files collected during inquiry within learning environments were not being fully leveraged for generating performance assessments (Scalise et al., 2009). This was partly due to the fact that the development of science simulations and games had "outpaced their grounding in theory and assessment" (Quellmalz et al., 2009, p. 1). As noted in our literature review, additional problems have arisen because traditional statistical methods, namely, classical test theory and item response theory (Hambleton & Jones, 1993), are not adequate for assessing or modeling rich inquiry skills that are multidimensional and typically multistepped in nature (Williamson et al., 2006).

In our work here, we present a rigorous and nuanced solution to performance assessment using log files that applies both top-down and bottom up-coding. Specifically, our approach is top down in that we used theoretically grounded principles from cognitive science, prior data (Buckley et al., 2010), and the prior literature on students' inquiry difficulties in order to generate categories for our text replay tagging of log files. Text replay tagging can be thought of as a form

of protocol analysis (Ericsson & Simon, 1980) for log files. Our approach is bottom up in that we used a machine-learned educational data mining algorithm in order develop our assessment models. When educational data mining is appropriately guided, as it was here, it provides a rigorous means for analyzing process data (Rupp et al., 2010) about students' learning such that a rich description is yielded. Our approach, a combination of text replay tagging and educational data mining that, in turn, is used to develop a detector, is, to our knowledge, unique in its application to the assessment of science inquiry skills.

Summary of Findings

In our article, we have described (a) how our detectors were developed for our phase change microworld for one inquiry skill of interest, namely designing and conducting experiments (see also Gobert et al., 2012; Sao Pedro, Baker, Gobert, et al., 2013); (b) how the detectors were validated, in turn, to assess students' skill at designing and conducting experiments in another physical science topic, namely, density; and (c) which assessment metrics were generated when our assessment detector was applied to density log files. Furthermore, we have shown that our detectors replicate human judgment, corresponding with standards considered acceptable for medical diagnosis (Fan et al., 2006; Yanamandra et al., 2011), about whether students were demonstrating this skill or not, as indicated by high values on four goodness metrics. Furthermore, this detector, developed in one scientific domain, phase change, was shown to effectively assess science inquiry skill in a second scientific domain, density, *without* modification or retraining of the machine-learned algorithm. The validation of the detector on a second microworld is of particular importance, as it establishes that the detector is generalizable in its valid applicability to a second physical science domain.

Note that our data demonstrate that the detectors can identify inquiry skill effectively in multiple domains, *not* that inquiry skills themselves are domain general. Others are engaged in this research program (Klahr & Nigam, 2004; Kuhn et al., 1992). Demonstrating that inquiry skills are domain general would require considerably more varied data than are presented here (multiple contexts per student). The issue of the domain generality of inquiry skills, although interesting and important, is beyond the scope of this article. From an assessment standpoint, a detector that can be generalized in order to assess and predict inquiry skills in multiple domains is valuable in its own right, regardless of whether students learn these skills in a domain-general fashion. Later, in "Limitations and Caveats", we discuss the issue of model generalizability further.

This approach for performance assessment has great potential to make significant headway in the field of science education on a few fronts, namely, in terms of contributions to performance assessment and analytical methods for performance assessment, to assessment of the skill of designing and conducting

experiments specifically, and to intelligent tutoring systems for science. Each is briefly discussed in turn.

Contributions to Performance Assessment and Performance Assessment Methods

This work contributes to the growing cadre of performance assessments for science inquiry (cf. Clarke-Midura et al., 2012; Ketelhut & Dede, 2006; Ketelhut et al., 2013; Quellmalz et al., 2011). Performance assessments, such as ours shown here, have important benefits over more traditional assessment types. First, computerized assessments can offer standardization in terms of delivery better than pencil-and-paper assessments. Second, all scoring, once developed, can be done automatically, alleviating the need for humans to score data. In terms of benefits over multiple-choice item tests, performance assessments can provide an authentic context, which can elicit rich inquiry and reasoning typically not tapped in either paper-based or multiple-choice tests (Quellmalz, DeBarger, Haertel, & Kreikemeier, 2005). (For a full discussion of the benefits of virtual performance assessments over traditional assessments, see Clarke-Midura et al., 2011.)

This work also contributes to the relatively new area of methods for analyzing log data, which is important because computer technologies are now being used for both learning and assessment of complex skills, such as inquiry (Quellmalz & Pellegrino, 2009). Analyzing log data using the methods we have shown here permits a rigorous way to better understand students' learning processes as opposed to their learning products (Rupp et al., 2010), as well as provides the potential for the scalability of these types of assessments. Furthermore, the validation and cross-validation processes used in our approach can be seen as analogous in terms of the level of rigor to those used in traditional psychometric measures of reliability and validity. In brief, the next generation of assessments (Quellmalz, Timms, Buckley, et al., 2012; Quellmalz, Timms, Silberglitt, et al., 2012) and methods for rigorous assessment of their data discussed here are very useful for informing instruction, informing the design of future assessments for science inquiry skills (Mislevy et al., 2012; Quellmalz et al., 2011), and potentially informing education policy as well (Quellmalz & Pellegrino, 2009).

It is important to note that the design of our microworlds and activities, and our analytic techniques, also allow us to circumvent the validity versus reliability tension of rich learning environments (Gobert & Koedinger, 2011). Specifically, simulations, because they are more authentic than multiple-choice items, add validity to the assessments gathered; however, because these complex tasks typically require longer periods of time to complete, there may be less of each type of measure. Although this can be seen as reducing reliability, the estimates of skill given by this type of assessment provide considerably more information than

multiple-choice items. In addition, in our environment students conduct multiple inquiry cycles, providing considerable data on the inquiry skill of interest. This is an advance because there have been, in the past, difficulties with measuring inquiry because of the amount of data required for reliable measurement (Shavelson et al., 1999).

Contributions to the Assessment of the Inquiry Skill—Designing and Conducting Experiments

Students have many difficulties designing and conducting experiments; in particular, they tend to change too many variables in the same trial (Glaser et al., 1992; Harrison & Schunn, 2004; Kuhn, 2005a; McElhaney & Linn, 2008, 2010; Reimann, 1991; Schunn & Anderson, 1998, 1999; Shute & Glaser, 1990; Tsirgi, 1980), and they also do not necessarily conduct trials in lockstep fashion (Buckley et al., 2006). Here we have shown that educational data mining can assess students' skills at designing and conducting experiments and auto-score them in a manner that handles their complexity; specifically, we are able to differentiate whether a student demonstrates the design of unconfounded experimental trials, even when the student chooses *not* to conduct sequential trials. This is an advance over other assessments of the CVS, which counts trials as evidence of this skill *if and only if* the trials are collected sequentially (Chen & Klahr, 1999; Klahr & Nigam, 2004; McElhaney & Linn, 2008, 2010).

As previously stated, these other approaches may fail, in our opinion, to accurately assess skill in more open-ended environments in which students *can* and *tend to* use a wide range of data collection strategies; for example, a student may run a few trials to observe the microworld, then change one variable to begin to test his or her hypothesis, then run a few more trials, and then target the variable of interest as a contrasting trial to one run earlier. In this case, a student may understand how to design controlled experiments but is engaging in *valid* exploration. However, a scoring metric that only counts successive pairwise controlled experiments would yield a low estimate of skill, failing to catch this corner case. Our method can, by contrast, account for experimentation strategies that are not pairwise, which represents an advance in the assessment of this key science inquiry skill.

Contributions to Intelligent Tutoring Systems for Science

Our work on assessment presented here and our current work on developing a pedagogical agent based on our assessment algorithms (Sao Pedro et al., 2013a) also contributes to and builds on the prior cognitive tutor research in other well-defined domains. Briefly, a considerable amount of work has been done on the development of intelligent tutoring systems for well-defined domains. Some examples

include problem solving in mathematics (Koedinger & Corbett, 2006; Razzaq et al., 2005), physics (Gertner & VanLehn, 2000), genetics (Corbett, Kaufmann, MacLaren, Wagner, & Jones, 2010), and computer programming tasks (Corbett & Anderson, 1995; Kasurinen & Nikula, 2009). These environments and tasks all elicit students' responses and data for which evaluation is relatively straightforward, because the tasks are all well defined (e.g., $2 + 3 = 5$) or the paths or actions a student follows in the environment to solve a problem are finite and explicit. As a result of these domains being well defined, procedures such as model tracing (Anderson, 1993; Anderson, Boyle, Corbett, & Lewis, 1990) can be implemented to auto-score students' data and provide real-time feedback.

A major hurdle in both assessing and scaffolding students is the fact that science inquiry is inherently ill defined, complex, and multifaceted, and the ways in which one can engage in inquiry are highly varied (Buckley et al., 2010; de Jong, & van Joolingen, 1998). Even though many specific subskills and proficiencies regarding inquiry have been identified (Zimmerman, 2007) for science inquiry, relatively little has been done with the goal of developing an intelligent tutoring system for scientific inquiry, with the exception of an earlier white paper that outlines how students might use a simulation, select tools for representing data, and then use evidence mapping to express their findings (Koedinger, Suthers, & Forbus, 1999). The inherent difficulty lies in the ill-defined nature of scientific inquiry, in which many if not all tasks are ill defined and thus their assessment is both ambiguous and open to interpretation (Duschl, 2003; Lynch, Ashley, Pinkwart, & Aleven, 2009; Voss, 2006). Our work presented here, which included breaking down the scientific inquiry skills of interest into subskills and then developing assessment detectors that can auto-score and eventually auto-scaffold students' inquiry, represents a significant advance toward these goals.

Limitations and Caveats

It is worth noting that the goodness of our detector was limited by the degree to which the students in this sample had representative inquiry behaviors relative to the eventual samples for which it would be used. For example, if students in urban schools or schools in another region conduct inquiry differently (perhaps because of different teaching practices), the detectors may not generalize well to these other students. Additional research is under way in which we are validating our detectors with other populations of students in order to study their generalizability.

Future Work

We are currently leveraging data mining for other inquiry skills, particularly complex inquiry skills such as interpreting data and warranting claims with data (NRC, 1996). This will involve similar methods to those described for the inquiry skill

of designing and conducting experiments, namely using text replay tagging using findings from the prior literature to develop our coding categories and then using machine learning to discover an educational data mining algorithm to identify such skills in students' log files.

Moreover, we are applying these methods to other science domains in life and earth science. A key contrasting feature between many physical science domains and those in life and earth science is that in the latter two domains, there are often multiple independent variables and/or dependent variables. For example, in both ecosystems and cell functions there are a number of interconnected, nonlinear elements that are interacting in a complex causal system (Jacobson & Wilensky, 2006; Yoon, 2008). One important implication of interacting independent variables is that the designing and conducting experiments is more complex because the classic CVS (cf. Chen & Klahr, 1999) is no longer effective, and thus more complex experimental strategies are needed. This may require us to develop new detectors, and one of our current projects addresses scientific inquiry skills in the context of the added complexity for both life and earth science topics in middle school.

Lastly, in our new work, we have now developed some adaptive scaffolds so that we can react in real time as students are conducting experiments in our virtual microworlds (Sao Pedro, 2013). Specifically, assessment skills detectors are being integrated into our learning environment in order drive real-time scaffolding of students for our physical science microworlds (Brusilovsky & Peylo, 2003; Sao Pedro et al., 2013a). Briefly, this requires us to develop assessment metrics for the full suite of inquiry skills and subskills (NRC, 1996; for details on our work to model other inquiry skills and subskills, see Gobert et al. 2012). A large advantage to this endeavor is the scalability it will afford (Leeman-Munk et al., 2013), because, of course, real-time, one-on-one tutoring is impossible in a classroom context. In our environment (http://www.inq-its.org), assessment algorithms will be used to drive a pedagogical agent to provide students with real-time, individualized scaffolds based on educational data mining–based metrics of inquiry skills. This will eliminate the separation between learning activities and assessment activities, which is the long-range vision for learner-centered environments (Quellmalz & Pellegrino, 2009; Quellmalz, Timms, Buckley, et al., 2012) so that the potential of technology can be more fully realized in classrooms.

REFERENCES

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, *42*, 7–49.

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, *18*(3), 287–314.

Baker, R., Corbett, A., & Wagner, A. (2006). Human classification of low-fidelity replays of student actions. In C. Heiner, R. Baker, & K. Yacef (Eds.), *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems* (pp. 29–36). Jhongli, Taiwan: Springer.

Baker, R., & de Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. In R. S. Baker, T. Barnes, & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38–47). Montreal, Canada.

Baker, R. S., Mitrovic, A., & Mathews, M. (2010). Detecting gaming the system in constraint-based tutors. In P. De Bra, P. Kobsa, & D. Chin (Eds.), *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization, UMAP 2010. LNCS 6075* (pp. 267–278). New York, NY: Springer-Verlag.

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.

Beck, J. E., & Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 353–362). Berlin, Germany: Springer.

Behrens, J. T. (2013, April). *Harnessing the currents of the digital ocean*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Ben-David, A. (2008). About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence*, *21*, 874–882.

Black, P. (1999). *Testing: Friend or foe? Theory and practice of assessment and testing*. New York, NY: Falmer Press.

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

Buckley, B. C., Gobert, J. D., & Horwitz, P. (2006). Using log files to track students' model-based inquiry. In *Proceedings of the 7th International Conference on Learning Sciences, ICLS* (pp. 57–63). Mahwah, NJ: Erlbaum.

Buckley, B., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking inside the black box: Assessments and decision-making in biologica. *International Journal of Learning Technology*, *5*(2), 166–190.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching*, *46*, 865–883.

Cen, H., Koedinger, K. R., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference of Intelligent Tutoring Systems* (pp. 796–798). Berlin, Germany: Springer.

Cetintas, S., Si, L., Xin, Y., & Hord, C. (2010). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies*, *3*(3), 228–236.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098–1120.

Chi, M., VanLehn, K., & Litman, D. (2010). Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems: 10th International Conference, ITS 2010* (pp. 184–193). Heidelberg, Germany: Springer.

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* (pp. 138–164). New York, NY: IGI.

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

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

Corbett, A., Kaufmann, L., MacLaren, B., Wagner, A., & Jones, E. (2010). A cognitive tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research, 42,* 219–239.

DeBoer, G. (1991). *A history of ideas in science education: Implications for practice*. New York, NY: Teachers College Press.

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* (pp. 231–252). Arlington, VA: NSTA Press.

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.

Desmarais, M. C., Meshkinfam, P., & Gagnon, M. (2006). Learned student models with item to item knowledge structures. *User Modeling and User-Adapted Interaction*, *16*, 403–434.

Dewey, J. (1910, January 28). Science as subject-matter and as method. *Science*, *31*, 121–127.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education.* New York, NY: MacMillan.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz (Ed.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age.

Duschl, R. A. (2003). Assessment of inquiry. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41–59). Arlington, VA: NSTA Press.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*(1), 36–48.

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

Fadel, C., Honey, M., & Pasnick, S. (2007). Assessment in the age of innovation. *Education Week*, *26*(38), 34–40.

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, *8*(1), 19–20.

Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction*, *30*(3), 207–245.

Fu, A. C., Raizen, S., & Shavelson, R. J. (2009, December 18). The nation's report card: A vision of large-scale science assessment. *Science*, *326*, 1637–1638.

Gertner, A. S., & VanLehn, K. (2000). Andes: A coached problem solving environment for physics. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, LNCS 1839* (pp. 133–142). Montreal, Canada: Springer-Verlag.

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

Gobert, J. (2005). Leveraging technology and cognitive theory on visualization to promote students' science learning and literacy. In J. Gilbert (Ed.), *Visualization in science education* (pp. 73–90). Dordrecht, The Netherlands: Springer-Verlag.

Gobert, J. (in press). Microworlds. In R. Gunstone (Ed.), *Encyclopedia of science education*. Heidelberg, Germany: Springer-Verlag.

Gobert, J., & Koedinger, K. (2011, September). *Using model-tracing to conduct performance assessment of students' inquiry skills within a microworld*. Presentation at the Society for Research Effectiveness, Washington, DC.

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*, *4*(1), 153–185.

Gobert, J., Sao Pedro, M., Toto, E., Montalvo, O., & Baker, R. (2011, April). *Science ASSISTments: Assessing and scaffolding students' inquiry skills in real time.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. *International Journal of Science Education, 22,* 891–894.

Haertel, G., Lash, A., Javitz, H., & Quellmalz, E. (2006, April). *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.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12(3),* 38–47.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

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

Jacobson, M. J., & Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. *Journal of the Learning Sciences*, *15*(1), 11–34.

Kasurinen, J., & Nikula, U. (2009). Estimating programming knowledge with Bayesian knowledge tracing. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education* (pp. 313–317). New York, NY: ACM Press.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, *70*(1), 163–173.

Ketelhut, D. J., & Dede, C. (2006, April). *Assessing inquiry learning*. Paper presented at the National Association for Research in Science Teaching, San Francisco, CA.

Ketelhut, D., Nelson, B., Sil, A., & Yates, A. (2013, April/May). *Discovering what students know through data mining their problem-solving actions within the immersive virtual environment, SAVE Science*. Presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). New York, NY: Cambridge University Press.

Koedinger, K. R., & MacLaren, B. A. (2002). *Developing a pedagogical domain theory of early algebra problem solving* (CMU-HCII Tech. Rep. No. 02-100). Pittsburgh, PA: Carnegie Mellon University.

Koedinger, K. R., Suthers, D. D., & Forbus, K. D. (1999). Component-based construction of a science learning space. *International Journal of Artificial Intelligence in Education*, *10*, 292–313.

Koomen, M. (2006, April). *The development and implementation of a computer-based assessment of science literacy in PISA 2006*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Kuhn, D. (2005b). What needs to be mastered in mastery of scientific method? *Psychological Science*, *16*, 873–874.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*, 495–523.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction*, *46*, 512–559.

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

Lee, D. M. C., Rodrigo, M. M. T., Baker, R. S. J. d., Sugay, J. O., & Coronel, A. (2011). Exploring the relationship between novice programmer confusion and achievement. In S. K. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Proceedings of the 4th Bi-Annual International Conference on Affective Computing and Intelligent Interaction*, *Part 1* (pp. 175–184). Berlin, Germany: Springer.

Leeman-Munk, S., Wiebe, E., & Lester, J. (2013, April). *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.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. (2009). Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, *19*(3), 253–266.

Massachusetts Department of Education. (2006). *Massachusetts science and technology/engineering curriculum framework*. Malden, MA: Author.

McElhaney, K., & Linn, M. (2008). Impacts of students' experimentation using a dynamic visualization on their understanding of motion. In P. A. Kirschner, J. J. G. van Merriënboer, & T. de Jong (Eds.), *Proceedings of the 8th International Conference of the Learning Sciences* (Vol. 2, pp. 51–58). Utrecht, The Netherlands: International Society of the Learning Sciences.

McElhaney, K., & Linn, M. (2010). Helping students make controlled experiments more informative. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010): Vol. 1. Full papers* (pp. 786–793). Chicago, IL: International Society of the Learning Sciences.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 935–940). Philadelphia, PA: ACM Press.

Mislevy, R. J., Behrens, J. T., DiCerbo, K., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*, *4*, 11–48.

Mislevy, R., Hamel, L., Fried, R., G., Gaffney, T., Haertel, G., Hafter, A., . . . Wenk, A. (2003). *Design patterns for assessing science inquiry (PADI Technical Report 1)*. Menlo Park, CA: SRI International.

Montalvo, O., Baker, R. S., Sao Pedro, M. A., Nakama, A., & Gobert, J. D. (2010). Identifying students' inquiry planning using machine learning. In R. Baker, A. Merceron, & P. Pavlik (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 141–150). Pittsburgh, PA: EDM.

National Assessment Governing Board. (2011). *Science framework for the 2011 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academies Press.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

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.

Organisation for Economic Co-operation and Development. (2010). *PISA computer-based assessment of student skills in science*. Retrieved from http://browse.oecdbookshop.org/oecd/pdfs/free/9810041E.pdf

Papert, S. (1980). Computer-based microworlds as incubators for powerful ideas. In R. Taylor (Ed.), *The computer in the school: Tutor, tool, tutee* (pp. 203–201). New York, NY: Teachers College Press.

Pardos, Z., Baker, R., Gowda, S., & Heffernan, N. (2011). The sum is greater than the parts: Ensembling models of student knowledge in educational software. *SIGKDD Explorations*, *13*(2), 37–44.

Pavlik, P., Cen, H., & Koedinger, J. (2009). Performance factors analysis—A new alternative to knowledge tracing. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 531–540). Brighton, England: IOS Press.

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

Quellmalz, E., DeBarger, A., Haertel, G., & Kreikemeier, P. (2005). *Validities of science inquiry assessments: Final report*. Menlo Park, CA: SRI International.

Quellmalz, E. S., DeBarger, A. H., Haertel, G., Schank, P., Buckley, B., Gobert, J., . . . Ayala, C. (2008). Exploring the role of technology-based simulations in science assessment: The Calipers Project. In *Science assessment: Research and practical approaches*. Washington, DC: NSTA.

Quellmalz, E., & Haertel, G. (2004). *Technology supports for state science assessment systems*. Washington, DC: National Research Council.

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., & Haertel, G. (2007, April). *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. S., & Pellegrino, J. W. (2009, January 2). Technology and testing. *Science*, *323*, 75–79.

Quellmalz, E. S., Silberglitt, M. D., & Timms, M. J. (2011). *How can simulations be components of balanced state science assessment systems* (Policy Brief). San Francisco, CA: WestEd. Retrieved from http://www.wested.org/cs/we/view/rs/1198

Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2012). 21st century dynamic assessment. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age.

Quellmalz, E., Timms, M., & Schneider, S. (2009). *Assessment of student learning in science simulations and games*. Washington, DC: National Research Council.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*, *49*, 363–393.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N., Koedinger, K. R., Junker, B., . . . Rasmussen, K. (2005). The Assistment Project: Blending assessment and assisting. In G. M. C. K. Looi (Ed.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 555–562). Amsterdam, The Netherlands: IOS Press.

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

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, *40*, 601–618.

Rowe, J., & Lester, J. (2010). Modeling user knowledge with dynamic Bayesian networks in interactive narrative environments. In C. G. Youngblood & V. Bulitko (Eds.), *Proceedings of the 6th Annual AI and Interactive Digital Entertainment Conference* (pp. 57–62). Palo Alto, CA: AAAI Press.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, *8*(4). Retrieved from http://escholarship.bc.edu/jtla/vol8/4

Rutherford, F. J., & Ahlgren, A. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology*. Washington, DC: American Association for the Advancement of Science.

Sao Pedro, M. A. (2013). *Real-time assessment, prediction, and scaffolding of middle school students' data collection skills within physical science simulations* (Unpublished doctoral dissertation). Worcester Polytechnic Institute, Worcester, MA.

Sao Pedro, M., Baker, R., & Gobert, J. (2013a). Incorporating scaffolding and tutor context into Bayesian knowledge tracing to predict inquiry skill acquisition. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 185–192). Memphis, TN.

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* (pp. 190–194). Leuven, Belgium: ACM Press.

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2013). 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. 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 (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 181–190). Pittsburgh, PA.

Sao Pedro, M., Gobert, J., & Baker, R. (2012, April). *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, Vancouver, British Columbia, Canada.

Sao Pedro, M., Gobert, J., Heffernan, N., & Beck, J. (2009). Comparing pedagogical approaches for teaching the control of variables strategy. In N. A. Taatgen & H. vanRijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1294–1299). Austin, TX: Cognitive Science Society.

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 (Eds.), *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences, ICLS 2010: Vol. 1. Full papers* (pp. 1024–1031). Chicago, IL: International Society of the Learning Sciences.

Scalise, K., Timms, M., Clark, L., & Moorjani, A. (2009, April). *Student learning in science simulations: What makes a difference?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*, 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: Erlbaum.

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

Siler, S., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010). Predictors of transfer of experimental design skills in elementary and middle school children. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems, Part II, LNCS 6095* (pp. 198–208). Pittsburgh, PA: Springer.

Slotta, J. D., & Linn, M. C. (2009). *WISE science: Web-based inquiry in the classroom.* New York, NY: Teachers College Press.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19(1)*, 1–14.

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, *23*, 488–511.

Timms, M., Clements, D. H., Gobert, J., Ketelhut, D. J., Lester, J., Reese, D. D., & Wiebe, E. (2012). *New measurement paradigms*. Retrieved from http://works.bepress.com/michael_timms/36

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

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

van Joolingen, W., & de Jong, T. (1993). Exploring a domain through a computer simulation: Traversing variable and relation space with the help of a hypothesis scratchpad. In D. Towne, T. de Jong, & H. Spada (Eds.), *Simulation-based experiential learning* (pp. 191–206). Berlin, Germany: Springer-Verlag.

Vapnik V. (1995). *The nature of statistical learning theory.* New York, NY: Springer-Verlag.

Voss, J. F. (2006). Toulmin's model and the solving of ill-structured problems. In D. Hitchcock & B. Verheij (Eds.), *Arguing on the Toulmin model: New essays in argument analysis and evaluation* (pp. 303–311). Berlin, Germany: Springer.

Walonoski, J., & Heffernan, N. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In M. Ikeda, K. Ashlay, & T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006. LNCS 4053* (pp. 382–391). Johngli, Taiwan: Springer-Verlag.

Williamson, D., Mislevy, R., & Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Erlbaum.

Wilson, M., & Bertenthal, M. (Eds.). (2006). *Systems for state science assessment.* Washington, DC: National Academies Press.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco, CA: Morgan Kaufmann.

Yanamandra, K., Grudin, M. A., Casaite, V., Meskys, R., Forsgren, L., & Morozova-Roche, L. (2011). a-Synuclein reactive antibodies as diagnostic markers in blood sera of Parkinson's disease patients. *PlosONE*, 6(4), e18513. doi:10.1371/journal.pone.0018513

Yoon, S. (2008). Using memes and memetic processes to explain social and conceptual influences on student understanding about complex socio-scientific issues. *Journal of Research in Science Teaching*, *45*, 900–921.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27,* 172–223.

Zohar, A., & David, B. A. (2008). Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition and Learning*, *3*(1), 59–82.