The Future of Assessment: Measuring Science Reasoning and Inquiry Skills Using Simulations and Immersive Environments

Jodi L. Davenport, WestEd, Oakland, CA USA, jdavenp@wested.org
Edys S. Quellmalz, WestEd, Redwood City, CA USA, equellm@wested.org
Jody Clarke-Midura, Harvard University, Cambridge, MA USA, jec294@mail.harvard.edu
Chris Dede, Harvard University, Cambridge, MA USA, chris_dede@harvard.edu
Janice D. Gobert, Worcester Polytechnic Institute, Worcester, MA USA, jgobert@wpi.edu
Kenneth R. Koedinger, Carnegie Mellon University, Pittsburgh, PA USA, koedinger@cs.cmu.edu
Marty McCall, Smarter Balanced Assessment Consortium, Portland, OR, mccall.marty@gmail.com
Michael J. Timms, Australian Council for Educational Research, Melbourne, Australia, mtimms9@gmail.com

Abstract: Simulations and immersive environments provide innovative ways to measure students’ science reasoning and inquiry skills. These computer-based assessments allow for dynamic displays of science systems that expand how phenomena, information, and data can be represented; they also allow for interactivity that provides new ways for learners to demonstrate their knowledge and skills. A number of groups have been working to create and evaluate next-generation assessments that both evaluate students on 21st Century scientific skills and provide evidence models for making inferences about student proficiency. In this symposium, researchers who are currently developing and testing simulation-based and immersive assessments to meaningfully assess science content and inquiry skills will share findings from classroom-based studies of students using the assessments. The presentations will be followed by a discussion from James Pellegrino, an expert in assessment design.

Symposium Objectives
Simulations and immersive environments provide innovative ways to measure students’ science reasoning and inquiry skills. However, do assessments using these environments provide more information than traditional tests about student proficiency of science reasoning and inquiry skills? Whereas factual knowledge can be easily assessed with traditional paper-and-pencil tests, more complex scientific reasoning and inquiry skills (e.g., systems thinking, designing investigations, gathering evidence, explaining observations) are more difficult to measure with static assessments. Dynamic and interactive assessment designs expand how phenomena, information, and data can be represented and increase the number of ways learners can show their knowledge and skills. As the field starts to integrate technology-based assessments, the challenges for K-12 educators and assessment developers are to create assessment tasks that allow students to demonstrate 21st century scientific skills and to create evidence models for making inferences about student progress and proficiency. We want to move beyond simply putting multiple-choice questions about declarative knowledge online.

In this symposium, we bring together researchers who have developed simulation-based and immersive assessments to meaningfully assess science content and inquiry skills. The researchers will share both the designs of these technology-based science assessments using the conceptual framework of evidence-centered design and the findings from studies using the assessments with students. The goals of the current session are not only to describe next-generation assessments, but also to provide principled frameworks for their design, use, and evaluation.

Symposium Overview
The three papers in this session present research findings on innovative, computer-based assessments that are designed to measure scientific reasoning and inquiry skills. The papers report classroom-based findings that demonstrate novel ways to use student actions in open-ended environments to evaluate student proficiency.

In the first paper, Gobert and Koedinger present data from a study of the innovative, simulation-based learning environment, Science Assistments (www.scienceassistments.org). The Science Assistments platform records each “move” as students engage in inquiry practices. The system then uses model tracing to evaluate the actions that students performed to determine what students know about science inquiry. As students create hypotheses and design experiments, the system updates a model that estimates student proficiency on inquiry and reasoning skills. The Science Assistments system leverages the affordances of model-tracing algorithms to detect different patterns of student behaviors, including genuine discovery and confirmation bias.

In the second paper, Clarke-Midura, McCall, and Dede investigate the use of an immersive environment as a platform for assessing student inquiry and reasoning skills. As students move their avatars through a 3-D world, they are able to make observations, gather and analyze data, and draw conclusions. Clarke-Midura, McCall, and Dede found that the performance assessment using the immersive environment was
a reliable measure of inquiry and reasoning skills, though only a small number of students in the classroom studies demonstrated high-level inquiry skills.

Finally, in the third paper, Davenport, Quellmalz, and Timms report the results of an empirical study to determine whether active, animations and interactive, simulation-based assessments are better than static assessments at distinguishing factual knowledge of scientific principles from meaningful inquiry and reasoning skills (e.g., generating predictions from observations, designing experiments, and drawing conclusions). A pool of 1566 students participated in a within-subjects design in their science classrooms. Students took science assessments in each of three different modalities; static, active (using dynamic animations), and interactive (using simulations). The results suggest that the dynamic and interactive assessments were more effective than the static assessment (most similar to traditional, paper-based tests) at distinguishing declarative, factual knowledge from deeper scientific reasoning and inquiry skills.

After the presentations, James Pellegrino, an expert on technology and assessment, will provide a brief discussion and lead the question and answer session.

Significance

This symposium will bring together a variety of perspectives on the principled design and evaluation of next-generation assessments using simulation-based and immersive technologies. The symposium will create an opportunity for a broader discussion of the theoretical and practical considerations for leveraging emerging technologies to meaningfully assess complex science knowledge and inquiry practices.

Using Model-tracing to Conduct Performance Assessment of Students’ Science Inquiry Skill at Conducting Experiments Within a Microworld


Introduction

Many national frameworks for science emphasize inquiry skills (e.g., NRC, 1996). However, in typical classroom practice, science instruction often focuses on rote learning in part because science process skills are difficult to assess (Fadel, Honey, & Pasnick, 2007) and rote knowledge is prioritized on high-stakes tests. Short answer assessments of inquiry have been used (cf., Alonzo & Aschbacher, 2004; Songer, 2006), however, these tend to not align well to current national frameworks (Quellmalz, Kreikemeier, DeBarger, & Haertel, 2006) and it is unclear whether they properly identify inquiry skills (Black, 1999; Pellegrino, 2001). Hands-on performance assessments are more authentic (Baxter and Shavelson 1994; Ruiz-Primo & Shavelson, 1996), however, these are seldom used in schools because of difficulty with reliable administration and the resulting high cost.

Science Assistments (www.scienceassistments.org) learning environment assists and assesses (hence, “assistments”) middle school students on inquiry so teachers can assess their students’ skills during instruction--in the context in which they are developing (Mislevy et al, 2002).

Framework

As a proof of concept for automated assessment of scientific inquiry skills, we used model-tracing (Corbett & Anderson, 1995; Koedinger & Corbett, 2006) to develop a cognitive model of science inquiry skills, particularly, the control for variables strategy (Chen & Klahr, 1999) and warranting claims with data. This model provides a rich qualitative, process-oriented scoring of students’ inquiry “moves” within a guided scientific inquiry simulation for the domain of state change. We address the validity of this automated approach to performance assessment both quantitatively, in terms of reliability and predictive validity, and qualitatively, in terms of providing rich traces of student inquiry steps and “mis-steps” or haphazard inquiry (Buckley, Gobert et al, 2010). Additionally, we present Cronbach’s alphas as reliability measures for each of our variables, and correlations with other inquiry tasks as additional construct validity data.

Methods and Data Sources

Participants. 78 eighth grade students (aged 12-14 years) from a public school in Central MA participated. Students belonged to one of six class sections and had one of two science teachers.

Materials. Pre- and post-tests for inquiry skills (n=12) and domain knowledge (n=7) were used. A Phase Change Microworld activity was used with which students engaged in a series of inquiry tasks.

Data Collection and Analysis: By applying model-tracing to students’ log data from their interactions with microworlds, we use production rules to code for: 1) CVS-relevance for each of the trials (using the control for variables strategy and relevant to the student’s hypothesis), 2) tested-and-true hypotheses for each of the trials
Our model tracer tracked whether: students’ initial hypotheses were scientifically accurate, whether the experimental trials they ran were relevant to their hypotheses, whether their trials used the control for variables strategy (Chen & Klahr, 1999), whether their final analysis entered was supported or unsupported by their data, and whether they had collected appropriate experimental evidence that supported their final conclusion (relevant controlled trials). Using data from the model-tracer, we calculated Cronbach’s alpha for our variables to ascertain the reliability across the 4 trials on each of the measures. The Cronbach’s alpha for the 4 CVS-relevant scores was 0.683; the Cronbach’s alpha for the 4 true-tested scores was 0.741; and lastly, the Cronbach’s alpha for the aggregate of the 2 inquiry scores across the 4 trials, %CVS+true-tested, was 0.774, indicating a high degree of internal consistency for each of the three measures.

Correlations were calculated between our auto-scored performance measures of inquiry with specific post-test inquiry items that should, in theory, be related. We obtained moderate correlations between our performance measures of inquiry and our post-test items for identifying an independent variable, identifying a dependent variable, and demonstrating the control of variables strategy (CVS).

Findings
In this paper we have shown that we can use model-tracing as a method of performance assessment for science inquiry skills, an ill defined domain. This builds upon the extensive work that has been done to date for well-defined domains such as math (Corbett & Anderson, 1995; Koedinger & Corbett, 2006). Additionally: 1) the reliability of our machine-scored measures of inquiry are highly consistent across the 4 Assistment activities or “trials,” suggesting that we can reliably capture students’ inquiry performance on these rich inquiry tasks, and 2) our measures are moderately correlated with post-test measures of inquiry performance for analogous concepts. Lastly, our data show that model-tracing can detect interesting patterns of student inquiry such as confirmation bias and overcoming confirmation basis. These are important data with respect to demonstrating auto-scoring of rich inquiry behaviors, but are also important, particularly the former, in terms of its implications for adaptive scaffolding of student inquiry, such as that being done by the Science Assistments group (www.scienceassistments.org; Gobert et al, 2007, 2009; Sao Pedro et al, in press).

Significance
This work makes contribution to theoretical understanding of scientific inquiry, to its assessment, and to technical methods to auto-score inquiry. This represents an advance in this area since to date there has been difficulty in separating inquiry from the domain-specific context in which it was learned (Mislevy et al., 2002; Gobert, Pallant, & Daniels, 2010), and difficulty measuring inquiry skills due to their complexity and the amount of data required for reliable measurement (Shavelson et al, 1999).

Assessing Science Inquiry using Immersive Virtual Environments
Jody Clarke-Midura, Marty McCall, & Chris Dede, Harvard Graduate School of Education

Introduction
Scientific inquiry is the method by which scientists interact with and study the world. While detailed definitions of inquiry can be complex, at its core the process is hypothesized to involve theorizing and investigating. For example, Kuhn and colleagues define inquiry learning as investigations where students individually or collectively investigate a set of phenomena (virtual or real) and draw conclusions about it (Kuhn, Black, Keselman, & Kaplan, 2000). Similarly, White, Collins, & Frederiksen (2011) offer a definition of inquiry as a process that oscillates between theory and evidence within the practice of argumentation. Given the multifaceted and open-ended nature of inquiry, it is not surprising that research has found existing methods for assessing science inquiry learning to be limited (Quellmalz, Kreikemeier, DeBarger, & Haertel, 2006; USDE, 2010).

Traditional science assessments either fail to align with the active nature of inquiry (Quellmalz et al. 2006) or have difficulty distinguishing what part of the complex reasoning process students do not understand (Gotwals & Songer, 2010). This is partly due to the difficulty of capturing students’ actions and behaviors as they perform a paper-based or hands-on task. Digital media, such as virtual environments and simulations, allow us to create tasks that are more characteristic of how students engage with inquiry in the real world; these processes and trajectories are unobtrusively captured as log data (Clarke, 2009). In doing so, these technologies allow us to create observations of student learning not possible via hands-on, paper-based, and multiple choice assessments (Clarke-Midura & Dede, 2010). In this paper, we discuss how virtual performance assessments can provide reliable observations of students’ inquiry knowledge. These virtual performance assessments are delivered via
an immersive environment and require students to solve an authentic scientific problem in a virtual context (see vpa.gse.harvard.edu).

**Framework**

We used the Evidence Centered Design framework (ECD; Mislevy, Steinberg, & Almond, 2003) to develop our assessments. ECD “provides a framework for developing assessment tasks that elicit evidence (scores) that bears directly on the claims that one wants to make about what a student knows and can do” (Shute, et al., 2007, p. 6). Using this framework, we reframed our science inquiry constructs into specific knowledge, skills and abilities (KSAs) aligned with national frameworks. We then converted the KSAs into proficiencies: data gathering, evaluating evidence, experimenting, and reasoning from evidence. Through the process of articulating the exact details of what is being measured and how it is being measured, it is easy to link the KSAs to evidence of student learning. Linking KSAs like this provides a high degree of validity that research has found often lacking in performance assessments (e.g. Linn, Baker, & Dunbar, 1991).

**Methods and Data Sources**

In the virtual performance assessments, students have the ability to walk around the virtual environment, make observations, gather data, and run tests in a laboratory, in order to solve a scientific problem (see figures 1 and 2 below). They then build a claim and support it with evidence they gathered.

![Figures 1 and 2: Screenshots of the Virtual Assessments.](image)

As mentioned above, we compiled our inquiry skills into proficiencies: data gathering, evaluating evidence, experimenting, and reasoning from evidence (making claims). Observations of students demonstrating these skills as they participated in the assessment were compiled into groupings. Unlike traditional tests that characterize student performance in terms of response correctness, the interactive environment records whether or not students engaged in an activity. Students have varying levels of understanding about the scientific process. They may, for example, know that scientists collect evidence and make observations, but may not understand the role of control data or how to interpret observations in the light of previous research. The pattern of engagement in the interactive environment and the examinees’ final demonstration of evidence and reasoning from this evidence provide information about understanding of the scientific process. In order to assess the reliability of these observations as evidence of a students’ proficiency, we asked the following research question: *Are observations of student’s skills (i.e. reasoning from evidence) behaving as if they are governed by a coherent skill?*

This study involved 643 middle school students in two states in the US (females=341). In order to answer the research question, we used Item Response Theory (IRT). IRT is a type of latent trait analysis that can be viewed as non-linear factor analysis. A trait is any skill or ability that determines the likelihood of a specific response to a test question. Observations generated from the log data were recoded from raw data as degrees of correctness (no, partial, or full credit for demonstrating the skill). Both the ability of the examinee and the difficulty of observation levels are on the same scale. Data were analyzed using WINSTEPS (Linacre, 2003) with a partial credit one-parameter model.

**Findings**

IRT analysis found the observations were providing evidence of a solid trait. Due to limited space, this paper presents results on students’ ability to “reason from evidence.” Results showed moderate fit statistics (Table 1). The item-total correlation coefficients are very high—they should be over .25—indicating that student who did well on the overall skill also did well on each observation. This is an indicator that observations are performing as expected by the model, supporting the claim that actions are governed by a coherent trait.
While we found that we were able to model latent traits based on observations of students actions in the world, there were surprising findings about how students were responding to the multivariate nature of the problem. Our findings were in line with Kuhn et al’s research (2000) on students’ misconceptions around multivariate systems:

- students had difficulty teasing out causal and non-causal factors on the outcome.
- students were not able to distinguish the additive effects that individual features contributed to the respective effects on the outcome.
- students focused on surface level features while solving the problem.

We are in the process of exploring various analytic approaches to modeling the log data that will provide further insight and diagnostic data on what actions lead to lower and higher inquiry abilities.

**Significance**

Science inquiry is a complex process. Our research shows that we can use virtual environments to simulate the complexity of inquiry, while reliably assessing student learning. We demonstrated the reliability of our measure using traditional methods (IRT). However, we are also exploring how we can model student observational data from the virtual assessment using Bayesian networks and cognitive diagnostic models. Only through deep understanding of how to measure and model inquiry will we better understand the best methods for teaching it and for preparing our students to understand the complexity of multivariate systems.

**Affordances of Dynamic and Interactive Assessments for Measuring Science Inquiry and Reasoning**

Jodi L. Davenport, Edys S. Quellmalz, WestEd, and Michael Timms, ACER

**Introduction**

What are the affordances of simulation-based assessments for eliciting science inquiry skills? Computer-based assessments can portray dynamic information and allow for simulations that are interactive and responsive to student input. Technology-based science tests have been piloted in international testing programs including the Programme for International Student Assessment (PISA) (NGB, 2006; Koomen, 2006). These interactive assessments appear to be ideally suited to measuring student proficiency on science inquiry and reasoning skills. However, computer-based assessments are substantially more costly to develop and have many technical requirements. In the current study, we investigate whether dynamic and interactive assessments are more effective than traditional, static assessments at discriminating student proficiency on three types of science practices: identifying scientific principles (e.g., stating or recognizing principles), using science principles (e.g., predicting or explaining), and conducting inquiry (e.g., designing experiments).

**Framework**

Our project uses the theoretical frameworks of evidence-centered design and model-based learning to identify the interconnected knowledge and skills that form the student models in our assessments. The evidence-centered assessment design framework shapes the development of the assessments in our study. The process begins with domain analysis, then specification of the student models (content and inquiry to be tested), task models (designs of assessment tasks and items) and evidence models (data for content and inquiry learning) (Mislevy et al., 2003). Research on model-based learning suggests that effective science learners form, use, evaluate, and revise their mental models of phenomena in a recursive process that results in more complete, accurate, and
useful mental models of a science system (Gobert & Buckley, 2000). Further, cognitive research shows that learners who internalize schema of complex system organization—structure, functions, and emergent behaviors—can transfer this heuristic understanding across systems (e.g., Goldstone & Wilensky, 2008).

The assessment tasks were designed to elicit evidence of science practices described in the 2009 National Assessment of Educational Progress (NGB, 2006). The assessments require students to carry out the science practices in the context of the content domain of ecosystems, and to explicitly consider the components, interactions, and emergent behaviors characteristic of all complex systems.

To ensure the construct validity of the items, we carried out expert reviews with 3 independent reviewers and student think-alouds with 10 middle school students. Both expert reviews and student think-alouds revealed that the items elicited the targeted science practices.

Methods and Data Sources

A sample of 1566 students from the classrooms of 22 teachers in 12 states in the United States participated in the study. Each student completed 3 versions of computer-based Ecosystems assessments that varied in the level of interactivity (static, active, and interactive). In the static version, students viewed still images. In the active version, students viewed animations, but did not conduct active investigations. In the interactive version, students designed and ran their own experiments. Items were aligned across the assessments to tap into the same science practice skills (identifying scientific principles, using science principles, and conducting inquiry) and subskills (e.g., predicting or explaining, drawing conclusions). See Figure 1 for examples of the different modalities of assessment. The assessments were given in three consecutive sessions and the order of the sessions was counterbalanced (e.g., some students receive the static versions during the first session and others receiving the interactive versions first). To analyze the results, we used a generalizability study (G-Study) analysis which estimates the correlations between items designed to tap into the three science practices, and can determine which sets of items appear to be measuring distinct constructs versus the same construct. Our hypothesis was that the interactive assessment would be the most effective at distinguishing student abilities on conducting inquiry from the other science practices, as only the interactive assessment allows students to actively engage in designing experiments and gathering data based on their designs. We also hypothesized that the active, dynamic assessment would be more effective than the static assessment at distinguishing between using science principle and identifying principles, as the active, dynamic assessment allows students to watch animations and make observations of dynamic systems to generate predictions and explanations.

Methods and Data Sources

A sample of 1566 students from the classrooms of 22 teachers in 12 states in the United States participated in the study. Each student completed 3 versions of computer-based Ecosystems assessments that varied in the level of interactivity (static, active, and interactive). In the static version, students viewed still images. In the active version, students viewed animations, but did not conduct active investigations. In the interactive version, students designed and ran their own experiments. Items were aligned across the assessments to tap into the same science practice skills (identifying scientific principles, using science principles, and conducting inquiry) and subskills (e.g., predicting or explaining, drawing conclusions). See Figure 1 for examples of the different modalities of assessment. The assessments were given in three consecutive sessions and the order of the sessions was counterbalanced (e.g., some students receive the static versions during the first session and others receiving the interactive versions first). To analyze the results, we used a generalizability study (G-Study) analysis which estimates the correlations between items designed to tap into the three science practices, and can determine which sets of items appear to be measuring distinct constructs versus the same construct. Our hypothesis was that the interactive assessment would be the most effective at distinguishing student abilities on conducting inquiry from the other science practices, as only the interactive assessment allows students to actively engage in designing experiments and gathering data based on their designs. We also hypothesized that the active, dynamic assessment would be more effective than the static assessment at distinguishing between using science principle and identifying principles, as the active, dynamic assessment allows students to watch animations and make observations of dynamic systems to generate predictions and explanations.

Findings
Consistent with our hypotheses, our G-study analysis found that the active, dynamic assessment was the most effective at distinguishing student performance between identifying science practices (e.g., declarative facts) and using science principles (e.g., making predictions and explanations). The interactive, simulation-based assessment was the most effective at distinguishing student performance on conducting inquiry from either of the other science practices. Table 1 shows the correlations between each of the science practices for each of the assessment modalities. Notice that the active, dynamic assessment had the lowest correlation between identifying and using principles, and that the interactive assessment had the lowest correlation between conducting inquiry and either identifying or using principles.

Table 1: G-study correlations between science practices across assessment modalities.

<table>
<thead>
<tr>
<th></th>
<th>Correlation between Identifying Principles and Using Principles</th>
<th>Correlation between Identifying Principles and Conducting Inquiry</th>
<th>Correlation between Using Principles and Conducting Inquiry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.92</td>
<td>0.80</td>
<td>0.91</td>
</tr>
<tr>
<td>Active/Dynamic</td>
<td><strong>0.80</strong></td>
<td>0.80</td>
<td>1</td>
</tr>
<tr>
<td>Interactive</td>
<td>0.82</td>
<td><strong>0.72</strong></td>
<td><strong>0.84</strong></td>
</tr>
</tbody>
</table>

**Significance**

The current study provides the first large-scale empirical evidence of the affordances of dynamic and interactive assessments for discriminating among science reasoning and inquiry skills. Dynamic assessments are more effective than static assessments at differentiating between factual knowledge and the ability to apply that knowledge in meaningful contexts. Simulation-based, interactive assessments, with environments that are responsive to student actions, were more effective than either static or active assessments at uniquely measuring students’ ability to conduct inquiry.

**References**


The research in the paper by Clarke-Midura and Dede was supported by grants from the Institute of Education Sciences (IES#R305A080141) and the Bill and Melinda Gates Foundation to Harvard University. Any findings, interpretations, and conclusions do not constitute an official position of the U. S. Department of Education. The authors gratefully acknowledge the contributions of Drs. Jillianne Code, Geordie Dukas, Michael Mayrath, Nick Zap, as well as the assistance of numerous students at the Harvard Graduate School of Education. The material in the paper by Davenport, Quellmalz, and Timms is based upon work supported by the National Science Foundation under Grant No. DRL-0814776 awarded to WestEd, Edys Quellmalz, Principal Investigator. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.