

Studying the Potential of Virtual Performance Assessments for Measuring Student Achievement in Science

Jody Clarke, Harvard University

Historically, researchers have shown the limitations of measuring complex cognition and inquiry processes via multiple-choice and constructed-response paper-and-pencil tests (Resnick & Resnick, 1992; Quellmalz & Haertel, 2004; NRC, 2006). These tests also demonstrate limited sensitivity to discrepancies between inquiry and non-inquiry based science instruction (Haertel, Lash, Javitz, & Quellmalz 2006). Attempts to address these limitations by designing hands-on and virtual performance assessments in late 1980's and 1990's encountered a number of technical, resource, and reliability problems in large-scale administration (Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999). Then, these problems were substantial enough to undercut the potentially greater construct validity for science inquiry that performance assessments can provide over paper-and-pencil tests.

However, virtual performance assessments based on modern interactive media could mitigate many of these historical problems, because today's technologies have advanced capabilities such as data-tracking, access to large data sets, GIS map visualizations, ability to contrast visualizations of different data, and model phenomena that can't be observed with the naked eye, that were not available a decade ago.

We are using a modified version of the Evidence-Centered Design framework to develop virtual performance assessments for scientific inquiry for use in school settings as a standardized component of an accountability program. This paper/poster briefly introduces our research and presents early results of our design process. Please note our grant is new and we are in early design stages of our first assessment.

Background

Quellmalz, Kreikemeier, Haydel-DeBarger, & Haertel (2007) report that, on paper-and-pencil tests such as the National Assessment of Educational Progress (NAEP), Third International Math and Science Study (TIMSS), and New Standards Science Reference Exams (NSSRE), inquiry is not measured effectively:

Some of inquiry abilities such as communication, research questions, and alternative explanations are tested barely or not at all. If these are valued standards, it would seem that some items should test them... It should be noted that the exams do not report sub-scores either for inquiry as a general standard or for separate inquiry abilities or components (page 27).

While some of these tests involve formats other than paper-and-pencil, the investigators also note that, "Even the hands-on performance tasks in these large-scale science tests are highly structured and relatively short (15-40 minutes), truncating the investigation strategies that can be measured" (page 1). Thus, despite the increasing focus of world-wide science standards on inquiry, paper-and-pencil assessments continue to demonstrate misalignment and validity issues in the measurement of this domain.

In the late 1980s and 1990s educators attempted to use performance assessments in accountability programs. However, the developers of both hands-on and virtual performance assessments encountered a number of technical, resource, and reliability problems in large scale

administration (Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999). At that time, these problems were substantial enough to undercut the potentially greater construct validity for science inquiry that performance assessments can provide over paper-and-pencil tests. We believe that virtual performance assessments based on modern interactive media could mitigate many of these problems encountered historically. The research questions we are addressing in this project are:

RQ 1: Can we construct a virtual assessment that measures scientific inquiry, as defined by the NSES? What is the evidence that our assessments are designed to test NSES inquiry abilities?

RQ 2: Are these assessments reliable?

Design Framework

The goal of an assessment is to provide valid inferences related to particular expectations for students (Linn et al, 2002). The virtual performance assessments we are developing will assess science content and process areas outlined in the NSES science content standards. Specifically, we will design assessments that address NSES Content Standard A: Science as Inquiry. We believe that science cannot be understood as content separated from the process that created that content (National Research Council, 1996) and will therefore develop assessments that cover students’ ability to do scientific inquiry and understandings about scientific inquiry within the science domains covered in NSES standards on life science, specifically Populations and Ecosystems.

Over the past decade, researchers have made significant advances in methods of assessment design. Frameworks such as the Assessment Triangle (NRC, 2001) and Evidence-Centered Design (Mislevy et. al. 2003, Mislevy & Haertel, 2006) provide rigorous procedures for linking theories of learning and knowing to demonstrations to interpretation. To ensure maximum construct validity, we will use a modified version of the Evidence-Centered Design framework.

Evidence Centered Design is a comprehensive framework that contains four stages of design: domain analysis, domain modeling, conceptual assessment framework and compilation, and a four phase delivery architecture. Phases 1 and 2 focus on the purposes of the assessment, nature of knowing, and structures for observing and organizing knowledge. In Phase 3, assessment designers focus on the student model (what skills are being assessed), the evidence model (what behaviors/performances elicit the knowledge and skills being assessed), and the task model (situates that elicit the behaviors/evidence). These aspects of the design are inter-related. In the compilation phase, tasks are created. The purpose is to develop models for schema-based task authoring and developing protocols for fitting and estimation of psychometric models. Phase 4 of the delivery architecture, focuses on the presentation and scoring of the task.

Following is our interpretation of the ECD framework (Mislevy et al, 2003; Mislevy & Haertel, 2003) that we are using to develop our assessments.

Modified ECD framework	Description
I. Domain Analysis	Develop purpose for assessment. Compile research on teaching and assessing inquiry as defined by NSES. Examine relationships between standards and existing

	assessments. Develop definition of competence. Develop competence of understanding. Consult experts in the fields about our chosen definitions and definitions of inquiry and learning objectives.
II. Domain Modeling	Use information from the domain analysis to establish relationships among proficiencies, tasks, and evidence. Explore different approaches and develop high-level sketches that are consistent with what they have learned about the domain so far. Develop narrative descriptions of proficiencies of inquiry, ways of getting observations that evidence proficiency, and ways of arranging situations in which students provide evidence of targeted proficiencies. Create graphic representations and schema to convey these complex relationships, and develop prototypes.
III. Conceptual Assessment Framework <ul style="list-style-type: none"> • Student Model • Evidence Model • Task Model 	<ol style="list-style-type: none"> 1. <i>Student Model</i>: what we are measuring in terms of students proficiency. 2. <i>Task Model</i>: how will students performance be captured? 3. <i>Evidence Model</i>: Identify behaviors and performances that reveal knowledge and skill identified in the student model. This bridges the task and student model.
IV. Assessment Implementation	Develop models for evidence. Develop statistical assembly and strategies and algorithms for test construction. Develop data structures and processes for implementing assessments. Develop back end architecture that will capture and score student data.
V. Assessment Delivery	Pilot assessments and evaluate results.

Research Methods

Validity is a central issue in test construction. According to Messick, “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. . .” (cited in Messick, 1994 p.1). In order to provide evidence that our assessment questions test students’ ability to do inquiry as outlined by the NSES standards, we plan to conduct a series of validity studies that provide evidence on construct validity. We will employ similar methods to those carried out in the Validities of Science Inquiry Assessments (VSIA)(Quellmalz, Kreikemier , Haydel-DeBarger, Haertel, 2007). We will conduct both an alignment analyses and a cognitive analyses of our assessments because these methods provide valuable, separate sources of validity evidence (Quellmalz, 2007).

To assess the reliability of performance assessments, we plant to conduct a series of generalizability studies. Generalizability theory (g-theory) is a statistical theory that allows decision makers to study the dependability of behavioral measures and procedures (Shavelson & Webb, 1991). It is a commonly used technique for making decisions and drawing conclusions

about the dependability of performance assessments (Baxter, 1995; Baxter and Shavelson, 1994; Pine et al, 1993; Shavelson, Baxter, Pine, 1991; Rosenquist, Shavelson, Ruiz-Primo, 2000). G-theory was first introduced as a statistical theory by Cronbach, Gleser, Nanda, & Rajaratnam (1972) to extend the limitations of using classical test theory, which provides an estimate of a person's true score on a test, by allowing researchers to generalize about a persons' behavior in a defined universe from observed scores (Shavelson, Webb, & Rowley, 1989). Further, classical test theory can estimate only one source of error at a time (Cronbach et al, 1972; Shavelson & Webb, 1991) whereas in g-theory, multiple sources of error can be measured in a single analysis.

Conclusion

The assessments we are creating will complement rather than replace existing standardized measures by assessing skills not possible via paper-pencil and multiple-choice or via hands-on performance assessment. One of the advantages of developing virtual assessments is that they will alleviate the need for extensive training for administering tasks. It is difficult to standardize the administration paper-based performance assessments, and extensive training is required to administer the tasks. With virtual assessments, we can ensure standardization by delivering instruction automatically via the technology.

A second advantage is that virtual assessments alleviate the need for providing materials and kits for hands-on tasks. Everything will be inside the virtual environment. Third, these performance assessments will be easier to administer and will require very little, if any, training of teachers. Scoring will all be done behind the scenes—there will be no need for raters or training of raters. Fourth, virtual assessments would alleviate safety issues and inequity due to lack of resources.

References

- Baxter, G. P. (1995). Using computer simulations to assess hands-on science learning. *Journal of Science Education and Technology*, 4, 21–27.
- Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, 21, 279-298.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Cronbach, L. J., Linn, R. L., Brennan, R. L, & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Haertel, G. D., Lash, A., Javitz, H., & Quellmalz, E. (2006). An instructional sensitivity study of science inquiry items from three large-scale science examinations. Presented at AERA 2007.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). "Accountability systems: Implications of requirements of the "No Child Left Behind Act of 2001." *Educational Researcher*, 31(6), 3-26.

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

Mislevy, R., & Haertel, G. (2006). Implications of Evidence-Centered Design for Educational Testing (Draft PADI Technical Report 17). Menlo Park, CA: SRI International.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.

National Research Council. (2006). Systems for state science assessment. Washington, DC: The National Academies Press.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Pine, J., Baxter, G., & Shavelson, R. (1993). Assessments for hands-on elementary science curricula. *MSTA Journal*, 39(2), 3, 5-19.

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

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., & Haertel, G. (2007). A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards. Presented at AERA 2007.

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction*. Norwell, MA: Kluwer Academic Publishers, 37-75.

Rosenquist, A., Shavelson, R.J., & Ruiz-Primo, M.A. (2000). On the "exchangeability" of hands-on and computer simulation science performance assessments. National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, Graduate school of Education & Information Studies, UCLA, Los Angeles, CA. 90024-6511.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education* [Special Issue: R. Stiggins and B. Plake, Guest Editors], 4(4), 347-362.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, 36, 61-71.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, CA: Sage.