Paper #4
Abstract Title Page

Title: Exploring the Utility of a Virtual Performance Assessment

Author(s): Jody Clarke-Midura, Jillianne Code, Nick Zap, Chris Dede
Affiliations of authors: Virtual Assessment Research Group, Harvard Graduate School of Education, 50 Church Street, Suite 422, Cambridge, MA 02138
Author's emails: {jody_clarke, jilliane_code, nick_zap, chris_dede}@mail.harvard.edu
Contact email for paper: jody_clarke@mail.harvard.edu

Abstract Body

Limit 4 pages single spaced.

Background / Context:

Current large-scale testing practices for measuring inquiry do not provide valid inferences about students' inquiry learning (Baxter, Elder, & Glaser, 1996; National Research Council (NRC), 2006; Quellmalz, 1984; US Department of Education (USDE), 2010). Research has indicated that multiple-choice tests are not a good measure of higher-order and complex skills (Resnick & Resnick, 1992). For example, a study by Quellmalz and colleagues (2005) found that the inquiry items on three science reference exams—National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the New Standards Science Reference exam did not align with the National Science Education Standards (NSES) definition of inquiry. While researchers acknowledge that current practices are not capturing the type of active *minds-on* processes that constitute inquiry, previous attempts at defining alternate measures have been overturned.

In the 1990s there was a shift towards developing alternate assessments in science education that measured students' conceptual understanding and higher-level skills like problem solving (R. L. Linn, 1994). Studies were conducted to assess the reliability and construct validity of these performance assessments and also the feasibility (i.e., cost effectiveness and practicality) of using them large scale (R. L. Linn, 2000). Research supports that performance tasks are valuable both for aiding learning and for providing formative feedback to teachers about ongoing student attainment. However, when used as summative assessments, performance tasks were found to have issues around task sampling variability (Shavelson, Baxter, & Gao, 1993), occasion-sampling variability (Cronbach, Linn, Brennan, & Haertel, 1997), and validity (R. L. Linn, Baker, & Dunbar, 1991). In addition, performance assessments are not as cost-effective as multiple choice tests (Stecher & Klein, 1997). A series of studies conducted compared computer-simulated performance assessments to paper-based performance assessments (Baxter, 1995; Baxter & Shavelson, 1994; Pine, Baxter, & Shavelson, 1993; Rosenquist, Shavelson, & Ruiz-Primo, 2000; Shavelson, Baxter, & Pine, 1991). Findings from these studies suggest that hands-on and virtual investigations do not tap the same knowledge (Shavelson, et al., 1991), that prior knowledge and experience influence how students solve the problem (Shavelson, et al., 1991), and that issues of exchangeability was confounded with inconsistencies in performance (Rosenquist, et al., 2000). We believe these outcomes from the 1990s are largely due to the intrinsic constraints of paper-based measures, limited timelines in which to pilot and research the alternative assessments, and the limited capabilities of virtual performance assessments constrained by what computers and telecommunications could accomplish more than a decade ago.

Since the performance based assessment studies of the 1990s, three advances have taken place that potentially enable virtual performance assessments capable of validly measuring the full complexity of scientific inquiry: 1) advances in cognitive science, 2) advances in statistics and measurement, and 3) advances in information and communication technologies. To illustrate the power and potential of these new types of performance assessments, this proposed session will describes our current research on one such model: virtual performance assessments (VPAs); 3D immersive technologies that aim to situate students in an environment that promotes inquiry and sets the context for assessment. VPAs are immersive three-dimensional (3D) environments, either single or multi-user, where participants engage in virtual activities and experiences. Such immersive environments support student experimentation and scientific reasoning in a virtual context by allowing students the ability to walk around an environment, make observations, gather data, and solve a scientific problem in context. Each participant takes on the identity of an avatar, a virtual persona that can move around the 3D environment. VPAs enable the creation and measurement of authentic, situated performances that are characteristic of how students conduct inquiry (Bransford, Brown, & Cocking, 1999). In fact, recent reports such as the *Science Framework for the 2011 National Assessment of Educational Progress* (NAGB, 2010), *PISA 2006: Science Competencies for Tomorrow's World* (OECD, 2007), *National Educational Technology Plan* (NETP; USDE, 2010), and the National Research Council's report on *Learning Science Through Computer Games and Simulations* (NRC, 2011) all recognize the potential of using technology to assess science inquiry. One advantage is that VPas make possible the automated, invisible, and non-intrusive collection of students' actions and behaviors during the act of learning (Pellegrino, et al., 2001). Rather than rely on student responses to questions about their knowledge, VPAs enable the capture and assessment of inquiry *in situ*.

Purpose / Objective / Research Question / Focus of Study:
With funding from the Institute of Education Sciences (IES), the Virtual Performance Assessment project at the Harvard Graduate School of Education is developing and studying the feasibility of immersive virtual performance assessments (VPAs) to assess scientific inquiry of middle school students as a standardized component of an accountability program (see http://vpa.gse.harvard.edu). The purpose is to provide states with reliable and valid technology-based performance assessments linked to state and NSES academic standards around science content and inquiry practices. In order to ensure construct validity, we are using the Evidence Centered Design framework (ECD; Mislevy, Steinberg, & Almond, 2003) to design our assessments. ECD formalizes the procedures generally done by expert assessment developers (Shute, Hansen, & Almond, 2007). Using the ECD approach, an evidentiary assessment argument is formed that connects claims, evidence, and supporting rationales. 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). We are conducting a series of studies around reliability, validity, and usability of our VPAs. Due to limited space, this paper presents research findings on usability and utility. However, we will also present results of our validity and reliability studies whose methods are reported in other papers (e.g Clarke-Midura, Code, Zap, Dede, in press).

Traditional assessments often focus on individual test items and rely on student affirmation as a response that indicates knowledge. In our VPAs, we base the evaluation of student performance on measurements captured as in-world interactions. These interactions allow us to assess what students know and do not know about science inquiry and problem solving. The assessment utility of any VPA is guided by design assumptions of how the interactions in the VPA facilitate the demonstration of students' knowledge and skills (see Clarke-Midura, et al., in press). To examine our situated design assumptions about how the VPA facilitates the immersive assessment of science inquiry, we conducted an empirical investigation of student perceptions of the assessment utility of the VPA for this purpose. Evaluating the assessment utility in this context focuses the research on the types of skills the VPA enables students to demonstrate; providing additional evidence that the VPA is a valid assessment of science inquiry. Building on the work of Nokelainen (2006) on pedagogical usability we have developed the Meaningful Assessment of Learning Questionnaire for Virtual Environments (MALQ-VE). Items on this scale are loosely based on items from the Pedagogically Meaningful Learning Questionnaire (PMLQ; Nokelainen, 2006) and help to establish how well each of our VPAs enable learner control, engagement in activity, added value for learning, flexibility, feedback, and valuation of previous knowledge (Code, Clarke-Midura, Mayrath, Zap, & Dede, 2011). Thus, the research question addressed in this paper is: *What are student perceptions of the assessment utility of the VPA?*

Setting:
This study took place in 10 eighth grade classrooms in three middle schools in the Northeast.

Population / Participants / Subjects:
There were 260 students in our sample (125 females, 135 males).

Intervention / Program / Practice:
We adapted 20 items from the Pedagogically Meaningful Learning Questionnaire (PMLQ; Nokelainen, 2006) into a survey we call, MALQ-VE. This included student perceptions of the following components of VPAs: learner control, learner activity, added value, flexibility, feedback, and valuation of previous knowledge (Code, Clarke-Midura, Mayrath, Zap, & Dede, 2011).

Students first worked individually at a computer and took our VPA, *Save the Kelp!* Immediately after finishing the assessment, which took approximately 60 minutes, students used our online survey software to complete the survey.

Research Design:
All eighth grade students in three schools participated in the intervention. Students were asked to state their agreement with a series of items using a 5 point Likert scale from 1 = strongly disagree to 5 = strongly agree.

Data Collection and Analysis:
Data was collected electronically using our in-house survey software. A classical item analysis, including an exploratory factor analysis, was conducted to assess the unidimensionality of this scale. A classical approach was chosen for this analysis because of the small relative sample size (N = 260).

Findings / Results:
The results of a classical analysis for the MALQ-VE are presented in Table 1. The distribution of item correlations (CITC) was from -.13 to .58. Since the CITC for items 1, 8, 10, 12, and 16 are low (< .25) they are poorly discriminating. These items were removed from subsequent analyses.

<insert table 1 here>

Table 1. Classical Item Analysis of the MALQ-VE (N = 260, α = .80, CI95 = .76, .83)

An exploratory factor analysis (EFA) was used to assess latent dimensionality of the MALQ-VE since the original validation of the PMLQ was conducted with elementary students and was designed for evaluating Learning Management Systems. The factors were extracted using Varimax rotation with Kaiser Normalization. The EFA on this data set revealed a three-factor structure as reported in Table 2: learner flexibility and feedback ($\alpha = 0.77$, CI95 = .72, .81), learner control ($\alpha = 0.69$, CI95 = .62, .74), and learner activity ($\alpha = 0.59$, CI95 = .50, .67). The calculated internal consistency of the entire scale is $\alpha = 0.83$, CI95 = .80, .86, above the acceptable level of $\alpha0 > .70$ (Tabachnick & Fidell, 2006) however, learner control and learner activity scales could be improved. Items on each of these scales will be revised in future studies.

<insert table 2 here>

Table 2. Factor Pattern and Structure Matrices for the MALQ-VE ($\alpha = 0.83$, CI95 = .80, .86)

Based on each newly defined factor, a summary analysis reveals (Table 3) that students strongly agreed or agreed that our VPA enabled increased learner flexibility and feedback (52.2%), control (46.5%) and activity (60.6%). However, these results reveal that there is still room for improvement of the overall student experience in each of these areas. Thus, we have been re-designing our assessments to include more flexibility and feedback, control and activity. We are currently in the process of analyzing data on the utility, validity and reliability of these newly designed assessments and will share the results in our presentation.

<insert table 3>

Table 3. Summary of Student Responses by Factor

Conclusions:

Current assessment approaches are inadequate at assessing how students develop sophisticated science reasoning, a key 21st century skill. We are building off research from the 1990s, as well as recent research that suggests performance assessments provide better measures of science inquiry (Darling-Hammond & Adamson, 2010; Lane, 2010; Lane & Stone, 2006). The assessments we are creating will complement rather than replace existing standardized measures by assessing skills not possible via item-based paper-and-pencil tests or hands-on real-world performance assessments.

The goal of our assessments is to simulate authentic experiences that provide a context for students to make meaningful choices and demonstrate their inquiry abilities. In order to meet this goal, we must establish the reliability and validity of our assessments, as well as the utility of students' perception of the experience. In addition to the research presented here, we are conducting cognitive task analyses and building cognitive flowcharts of students' inquiry learning and problem solving strategies throughout the assessment. Through these pilots and studies, we hope to establish alternate methods of assessment that provide valid and reliable evidence of students' inquiry learning.

Appendices.

Appendix A. References

Baxter, G. P. (1995). Using computer simulations to assess hands-on science learning. *Journal of Science Education and Technology, 4*(1), 21-27.

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*(2), 133-140.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*(3), 279-298.

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

Code, J. , Clarke-Midura, J., Mayrath, M. & Dede, C. (2011). Student perceptions of the assessment utility of immersive virtual assessments. Paper submitted to the AERA 2011 Annual Meeting New Orleans, LA.

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*(3), 373-399.

Darling-Hammond, L., & Adamson, F. (2010). Performance Assessment: The State of the Art. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Lane, S. (2010). Performance Assessment: The State of the Art. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Lane, S., & Stone, C. A. (2006). Performance Assessments. In R. L. Brennan (Ed.), *Educational Meaurement*. Westport, CT: Praeger.

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.

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

National Assessment Governing Board (NAGB). (2010). Science Framework for the 2011 National Assessment of Educational Progres. Washington, D.C.

National Research Council (NRC). (1996). *National Science Education Standards*. Washington, DC: National Academie Press.

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

National Research Council (NRC). (2010). A Framework for Science Education: Preliminary Public Draft Retrieved December 12, 2010, from http://www7.nationalacademies.org/bose/Standards_Framework_Homepage.html

National Research Council (NRC). (2011). *Learning Science Through Computer Games and Simulations*. Washington, DC: The National Academies Press.

Nokelainen, P. (2006). An empirical assessment of pedagogical usability criteria for digital learning material with elementary school students. *Educational Technology & Society, 9*(2), 179-197.

Organisation for Economic Co-operation and Development (OECD). (2007). *PISA 2006: Science competencies for tomorrow's world. Volume 1: Analysis.* Paris: OECD.

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

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

Quellmalz, E. (1984). Successful Large-Scale Writing Assessment Programs: Where Are We Now and Where Do We Go from Here? *Educational Measurement: Issues and Practices, 3*(1), 29-35.

Quellmalz, E., Kreikmeier, P., DeBarger, A. H., & Haertel, G. (2005). *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, San Francisco, CA.

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* (pp. 37-75). Norwell, MA: Kluwer Academic Publishers.

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, M. A. (2000). *On the "exchangeability" of hands-on and computer simulation science performance assessments*. (CSE Technical Report 531). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347-362.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility*. (RR-07-26). Princeton, NJ: Educational Testing Service.

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.

US Department of Education (USDE). (2010). National Education Technology Plan 2010. Washington, DC: US Department of Education.

Appendix B. Tables and Figures

Table 1. Classical Item Analysis of the MALQ-VE (N = 260, α = .80, CI_{95} = .76, .83)

					Response Category[d]				
Item	M[a]	SD	CITC[b]	α[c]	1	2	3	4	5
1	3.06	1.09	**-0.13**	0.82	16	70	84	62	28
2	3.48	0.98	0.48	0.78	10	31	74	115	30
3	3.54	0.92	0.48	0.78	9	21	80	120	30
4	3.43	0.96	0.38	0.79	8	38	74	115	25
5	3.26	1.07	0.53	0.78	10	63	67	89	31
6	3.01	1.16	0.33	0.79	30	62	64	84	20
7	3.15	1.24	0.44	0.78	31	46	79	60	44
8	2.77	1.03	**0.20**	0.80	24	91	78	55	12
9	3.40	1.03	0.45	0.78	12	42	64	113	29
10	3.05	1.08	**0.15**	0.80	21	63	76	82	18
11	3.75	1.11	0.33	0.79	13	28	37	114	68
12	3.62	0.92	**0.20**	0.79	4	29	68	121	38
13	3.12	1.06	0.58	0.77	16	62	79	82	21
14	3.32	1.00	0.42	0.78	16	31	88	103	22
15	3.49	0.90	0.47	0.78	5	31	83	114	27
16	3.36	0.96	**0.19**	0.80	8	40	87	100	25
17	2.90	1.27	0.48	0.78	48	55	59	72	26
18	3.56	0.91	0.57	0.78	6	30	62	136	26
19	3.59	0.88	0.37	0.79	7	20	73	133	27
20	3.58	1.02	0.36	0.79	6	37	66	103	48

Note: CITC = Corrected Item Total Correlation; Bolded items have a CITC < .25 and are poorly discriminating;
[a] Item mean is a classical test theory (CTT) indicator of difficulty. [b] Indicates item discrimination. [c] α if item is deleted; [d] 1 = strongly agree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree.

Table 2. Factor Pattern and Structure Matrices for the MALQ-VE ($\alpha = 0.83$, $CI_{95} = .80, .86$)

Item	Factor 1	Factor 2	Factor 3	h^2
Learner Flexibility and Feedback ($\alpha = 0.77$, $CI_{95} = .72, .81$)				
7	**0.62**	0.25		0.80
13	**0.60**	0.30	0.16	0.38
2	**0.58**	0.29		0.41
3	**0.56**	0.16	0.21	0.48
15	**0.52**		0.29	0.38
14	**0.42**	0.15	0.27	0.47
11	**0.37**	0.25	0.18	0.40
Learner Control ($\alpha = 0.69$, $CI_{95} = .62, .74$)				
5	0.30	**0.62**		0.21
9	0.20	**0.56**	0.13	0.49
4		**0.49**	0.17	0.38
17	0.36	**0.45**		0.43
6	0.10	**0.44**	0.15	0.38
Learner Activity ($\alpha = 0.59$, $CI_{95} = .50, .67$)				
20		0.24	**0.58**	0.27
18	0.28	0.35	**0.51**	0.39
19	0.23		**0.46**	0.36

Table 3. Summary of Student Responses by Factor

	Response Category[a]				
	1	2	3	4	5
Learner Flexibility & Feedback	5.5%	13.7%	28.6%	38.9%	13.3%
Learner Control	8.3%	20.0%	25.2%	36.4%	10.1%
Learner Activity	2.4%	11.2%	25.8%	47.7%	12.9%

Note: [a] 1 = strongly agree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree.