Formal Game-Based Assessments: The challenge and opportunity of building next generation assessments

Jody Clarke-Midura, Harvard Graduate School of Education, jec294@mail.harvard.edu
Jennifer Groff, Learning Games Network, jen@learninggamesnetwork.org

Abstract: National and global initiatives are starting to put pressure on testing systems and companies to change how learning is being measured. As a result, testing companies have begun turning to digital media as possible solutions for next generation assessments. While testing companies may be under pressure to change, the need for rethinking how we approach measuring learning involves a much greater shift than simply putting assessments "on line." It requires a framework that combines attributes from both game and assessment design. This paper will discuss the tensions between principles of good game-design and assessment design. We offer design insights and a suggested framework for designing/developing game-based assessments grounded in two case studies. Case 1 illustrates the tension between principles of good game design and what is required of assessments. Case 2 illustrates how some of the principles of good game design can actually be applied to assessment frameworks.

Introduction

> *One thing I never want to see happen is schools that are just teaching the test because then you're not learning about the world, you're not learning about different cultures, you're not learning about science, you're not learning about math. All you're learning about is how to fill out a little bubble on an exam and little tricks that you need to do in order to take a test and that's not going to make education interesting.*
>
> President Barack Obama, March 28, 2010

We are in a unique time where a confluence of events is creating the opportunity to re-think what it means to assess learning in the 21st century. National and global initiatives are starting to put pressure on testing systems and companies to change how learning is being measured (e.g. Race to the Top, Cisco, Intel and Microsoft's Assessment and Teaching of 21st Century Skills, development of new Common Core State Standards). As a result, testing companies have begun turning to digital media as possible solutions for next generation assessments. Many assessment organizations are creating pipelines to integrate digital game and new media designers and developers into this space—inviting digital media designers and learning scientists to sit on advisory boards, hiring game designers to explore the possibilities of using game-based assessments, and holding meetings on the future of assessment with a wide variety of stakeholders. However, while testing companies may be under pressure to change, the need for rethinking how we measure student learning involves a much greater shift than simply putting assessments "on line." It requires a framework that combines attributes from both game-design and assessment design. In this symposium, we will discuss the tensions between principles of good game-design (e.g. Gee, 2003; 2011) and assessment design (e.g. ECD, Mislevy, Steinburg, Almond, 2003). In doing so, we will offer design insights and a suggested framework for designing/developing game-based assessments that is grounded in two case studies: (1) the Learning Games Network's collaboration with ETS on developing game-based assessments and (2) efforts to design assessments using digital media in the Virtual Performance Assessment Project. In the following sections we briefly present the background context for our research, discuss an assessment framework, introduce the cases, and then conclude with the discussion of design principles.

Background

While there have been efforts to change standardized assessment programs in the past, they did not have the financial or policy-level support that is driving current initiatives. For example, efforts to use alternate assessment approaches such as performance-based measures for science (Linn, 1994), performance-based technology assessments (Baxter, 1995; Baxter and Shavelson, 1994; Pine, Baxter, Shavelson, 1993; Shavelson, Baxter, Pine, 1991; Rosenquist, Shavelson, Ruiz-Primo, 2000), and portfolios (Koretz, Stecher, Klein, & McCaffrey, 1994) were not robust enough to replace current

testing programs due to their inability to compete with the reliability, scalability, and cost-effectiveness of multiple-choice and open-response approaches (Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999). Two additional reasons for why previous efforts did not work are due to the fact that (1) the design teams were not comprised of teams containing: designers, psychologists, content specialists, and measurement specialists (either one or more were missing) and (2) not enough time was spent piloting items (NRC, 2010). While research has shown that multiple-choice and open-response tests are not good measures of higher-order thinking and cognitive skills (Resnick & Resnick, 1992; Quellmalz & Haertel, 2004), they have remained the default approach. However, advances in the learning sciences, technology/digital media, and measurement models over the past decade set the context for exciting opportunities for developing next generation assessments. We briefly discuss some of them below.

Data Capturing & Learning Analytics

Digital games and media allow for the capturing of data and observations of student learning that are not possible via multiple-choice and paper-based tests. As students interact with the digital environment, their actions can be captured via log data. These data can be utilized to explore learning trajectories, processes, and attempts at problem solving. Analysis of log data can provide more insight on learning by providing information on what led to incorrect as well as correct answers. Research on learning analytics is an emerging field and has gained increasing attention in the last several years, with initial endeavors in this space showing much promise (e.g. Baker, 2009; Roll, Aleven, Koedinger, 2010, Shute, 2011; Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2011; Shute, Masduki, Donmez, 2010). Next generation assessments will allow us to think differently about data and the kinds of algorithms used to model learning.

Assessment Design Frameworks

Over the past decade, researchers have made significant advances in methods of assessment design. Frameworks such as Evidence-Centered Design ([ECD] Mislevy et. al. 2003; Mislevy & Haertel, 2006) provide rigorous procedures for linking theories of learning and knowing to demonstrations to interpretation. ECD is a comprehensive framework that contains four stages of design: domain analysis, domain modeling, conceptual assessment framework and compilation, and 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 in-world interactions elicit the knowledge and skills being assessed), and the task model (situations 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. While the popularity of using ECD has increased, most projects developing assessments using digital media have adapted the framework to assess interactions and trajectories (e.g. Shute & Torres, 2011; Clarke-Midura, Code, Zap, & Dede, 2012, Rupp, Gushta, Mislevy, & Shaffer, 2010). As we will show in this paper, it is important to find a framework that incorporates affordances of both game design and assessment design.

Case 1: ETS and LGN

In the summer of 2011, the US Educational Testing Service (ETS) partnered with the Learning Games Network (LGN) to collaborate on the development of new testing modules that were designed from a digital games approach. The partnership was sought to purposefully experiment in this space and explore a new opportunity of developing testing scenarios. Ultimately, the collaboration surfaced a very real tension, and an innate opportunity:

> *There is a considerable and fundamental difference between learning games and assessment 'games'—yet much of this chasm is result of the vastly different paradigms these groups work from, and given the right space and support to explore this interdisciplinary work, real opportunity for innovation exists.*

From this work, it was very clear to us that this is most certainly an emerging space, and the entire field of assessment is just starting to figure out this nexus. We knew coming into the work there was a big difference between games and summative assessment, and we knew there would be some losses on both ends, but both groups acknowledged they greatly underestimated just how much that

challenge has been. The fundamental tension that emerged is that many games try to create a context that instills learning first, but that is not a goal in assessment—in fact, it's the exact opposite, formal assessment designers do not want the experience to instill anything because they want a baseline appraisal of the students knowledge and ability. Yet, we believe this tension does not mean there isn't an opportunity and possibility in this nexus. Rather, we see the challenge as being both groups coming from very different positions—which must be mitigated first in order to find the ripe opportunity in that nexus. Most certainly interdisciplinary work, our reflections align with suggestions from the research on successful interdisciplinary collaborations: both of these groups have found that when starting a new collaboration with non-traditional partners (Boix Mansilla, 2006), there is a phase of vocab reconciliation, where it takes time, space and the work itself to allow these two camps to get on the same page.

The ETS-LGN collaborated produced several fruitful artifacts based on various game designs and dynamics, including RPG models, which subsequently have been developed and adopted into national testing frameworks. However, more critically, it elucidated the critical areas to be worked through in order to achieve prosperous innovation at this nexus. What is needed is a focused workshop effort where interdisciplinary participants *create together* a shared vocabulary and understanding about each other's perspectives and needs in the design process.

Case 2: Virtual Assessment Project

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 goal of the research is to provide the field with working examples of reliable and valid technology-based performance assessments linked to state and national academic standards for science content and inquiry processes.

The virtual performance assessments are designed in the Unity game development engine (Unity Technologies, 2010). The immersive nature of the three-dimensional (3D) environment allows for the creation and measurement of authentic, situated performances that are characteristic of how students conduct inquiry (NRC, 2000). Students have the ability to walk around the environment, make observations, gather data, and solve a scientific problem in a context. Further, these environments enable the automated, invisible, and non-intrusive collection of students' actions and behaviors during the assessment play. These data allow us to build rich trajectories of student performance.

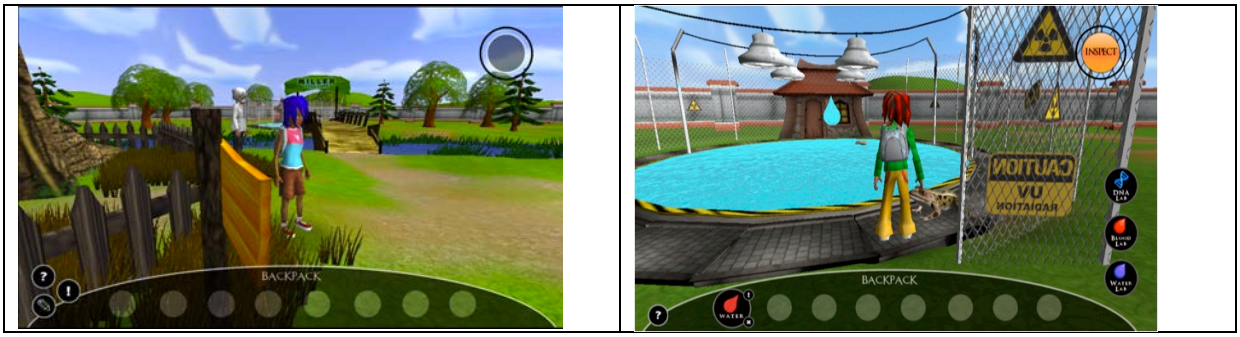
VPA Design Framework

In order to ensure that the assessments measure what we intended them to measure (inquiry), we used a modified version of the Evidence Centered Design (ECD) framework (Mislevey & Haertel, 2006; Mislevy & Rahman, 2009) to design the assessments (see Table 2 below). Using the ECD approach allowed us to articulate every aspect of the assessment from the knowledge, skills, and abilities (KSAs) that they are measuring to the types of evidence that will allow one to make claims about what students know. Using this framework, we have reframed science inquiry constructs (theorizing, questioning and hypothesizing, investigating, analyzing and synthesizing) into specific (KSAs) aligned with current national standards. 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 measure of validity that research has found often lacking in performance assessments (e.g. Linn, Baker, & Dunbar, 1991).

The skills we are measuring in this particular VPA focus on gathering data around a claim, making a claim, and supporting it with evidence and reasoning—skills that we argue are difficult to capture in multiple choice and open response tests. By setting up the assessment in a game-based environment, we can follow students' trajectories of data gathering. We then can correlate their interactions to the claims they build and the evidence and reasoning they use to support those assertions. Each challenge in our assessments relies on students collecting data and providing evidence to support a claim, and students' scores are based on the evidence and reasoning they provide for a given claim.

Traditional assessments often focus on individual test items and rely on student affirmation as a response that indicates knowledge. In the VPAs, the evaluation of student performance is based 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 series of interactions result in rich observations that enable us to make a fine distinction of students' understanding of the various facets of inquiry. Designing for interactions involves providing experiences for students that not only model how a scientist conducts science but also provide opportunities for learning and feedback on learning. Existing paper-based models for assessing science are not able to model the complexity of science practices and processes. See figures 1 and 2 below for images of the assessment.



Figures 1 & 2: **Screenshots of the Virtual Assessments.**

As seen in the images above, our assessment has the look and feel of a videogame, yet places students at the center of a scientific problem that they have to solve. Thus, our attempt is develop assessments that measure students' science learning *in situ*.

Next Generation Design Framework

Lessons learned from our cases are that the principles we apply to good game design involve play as learning and learning from play. However, when using games for assessment, play becomes performance and we need to think about performance as play and demonstration of learning (Clarke-Midura, in press). Various traits of effective learning games have been proposed in the literature (Osterweil & Klopfer, 2011; Gee, 2011; Gee, 2005, which we have synthesized into a list of dimension of characteristics of good learning games:

Dimensions of Characteristics of Learning Games
1. Freedom to Fail
2. Freedom to Experiment
3. Freedom of Identity
4. Freedom of Effort
5. Narrative
6. Agency
7. Interaction
8. Well-ordered problems
9. Clear goals
10. Copious feedback
11. Customization
12. Well-designed experiences
13. "Pleasantly Frustrating"
14. Mentoring in game and meta-game
15. Performance before competence
16. Cycle of Expertise
17. Smart tools
18. Non-linear learning
19. Distributed knowledge
20. Information just-in-time and on demand
21. Model-based and system thinking
22. Production and innovation

Table 1**: Dimensions of characteristics of good learning games.**

These traits often appear in good learning games, to varying degrees—which is why we emphasize the 'dimensional' aspect to these characteristics; however they are often observed in a good learning game as they embody a distinct pedagogical approach. The first case study, LGN and ETS, will illustrate the tension between what we agree are principles of good game design and what is required of assessments. In the second case study, VPA, we illustrate how some of the principles of good game design can actually be applied during the compilation phase of ECD, when designing the tasks and interactions. It is important to start with the knowledge, skills, and abilities you want to measure and then work backwards to come up with examples of evidence. How do you know that a student has demonstrated a particular knowledge or skill? ECD forces you to think through the kinds of performances or interactions that provide evidence that a student knows or understands a skill. We refer to this as performance as play. Take aspects of play and good game design and turn them into *performances of knowing*. Table 2 below presents the integration of Gee's principles with the ECD framework.

Modified ECD framework	Description
I. Domain Analysis	• Develop purpose for assessment. • Develop definition of competence. • Consult experts in the fields about our chosen definitions.
II. Domain Modeling	• Use information from the domain analysis to establish relationships among proficiencies, tasks, and evidence. • Develop high-level sketches that are consistent with what they have learned about the domain so far. • Create graphic representations and schema to convey these complex relationships, and develop prototypes.
III. Conceptual Assessment Framework: • Student Model • Observation/Tasks Model • Interpretation/ Evidence Model	*Student model:* What complex of knowledge, skills, or other abilities should be assessed. *Observations/Tasks:* Identify kinds of tasks or situations (interactions) that will prompt students to say, do, or create something that demonstrates important knowledge, skills, and competencies. *Evidence:* Identify behaviors and performances that reveal knowledge and skill identified in the student model. Identify and summarize evidence.
IV. Compilation: • Task creation • Statistical Assembly • Assessment Implementation	Develop tasks based on conceptual assessment framework that include characteristics of good game design. • Problem-solving • Clear goals • Well-designed experiences • Well-ordered problems • Smart tools • Information just-in-time and on demand • Model-based and system thinking Develop models for evidence. Develop statistical assembly and strategies and algorithms for test construction.
V. Four-Process Delivery Architecture: • Presentation • Response Scoring • Summary Scoring • Activity Selection	Develop data structures and processes for implementing assessments. Develop back-end architecture that will capture and score student data. Develop prototype. Pilot.
VI. Refinement	Refine assessment based on pilot data. Iterative cycle.

***Table 2*: Modified Evidence Centered Design Framework.**

Conclusion

The need and demand for richer forms of data about student learning and ability has never been greater; at the same time, never has the energy and evidence of the opportunity of learning games. Many state, national, and international testing companies are starting to transition to technology for delivering and administering assessments. While numerous initiatives exist for promoting change and innovation in current assessment systems, it is important that we learn from historical attempts at changing assessment. Digital media and game-based assessments offer potential to design innovative approaches for measuring learning and providing observations that are not possible with

multiple-choice and open-response tests. Next generation assessments require a collaborative team of designers with expertise in measurement, content, learning & cognition, and design of digital media/games.

References

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).

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.

Boix Mansilla, V. (2006). Assessing expert interdisciplinary work at the frontier: An empirical exploration. *Research Evaluation, 15*(1), 17-29.

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

Clarke-Midura, J. (in press). *Performance as play*.

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.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gee, J. (2011). Good video games and good learning. Unpublished essay, accessible at http://dmlcentral.net/resources/4578

Gee, J. (2005). Learning by design: Good video games as learning machines. *E-Learning and Digital Media, 2*(1), 5-16.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5-16.

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

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5-21.

Mislevy, R., & Haertel, G. (2006). Implications of evidence centered design for educational testing PADI Technical Report 17. Menlo Park, CA: SRI International.

Mislevy, R., & Rahman, T. (2009). Design pattern for assessing cause and effect reasoning in reading comprehension PADI Technical Report 20. Menlo Park, CA: SRI Interantional.

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

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: The National Academies Press

National Research Council. (2010). "Political Experiences and Considerations." State Assessment Systems: Exploring Best Practices and Innovations: Summary of Two Workshops. Washington, DC: The National Academies Press.

Osterweil, O., & Klopfer, E. (2011). Are games all child's play? In de Freitas, S. & Mahrg, P. (eds)., *Digital Games and Learning*. Continuum: London.

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

Quellmalz, E., & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323(5910), 75-79.

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.

Roll, I., Aleven, V., & Koedinger, K. R. (2010). The invention lab: Using a hybrid of model tracing and constraint- based modeling to offer intelligent support in inquiry environments. In V. Aleven, J. Kay, & J. Mostow (Eds.), Proceedings of the 10th International Conference on Intelligent Tutoring Systems (pp. 115-24). Berlin: Springer Verlag

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.

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).

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. 2011. Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*.

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

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments . *Technology, Instruction, Cognition, and Learning, 8*(2), 137-161.

Shute, V. J. & Torres, R. (2011). Where streams converge: Using evidence-centered design to assess Quest to Learn. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 91-124). Charlotte, NC: Information Age Publishing.

Unity Technologies. (2010). Unity 3D Engine. from http://unity3d.com/