Response to Assessment of Student Learning in Science Simulations and Games

John T. Behrens

Cisco

Jbehrens@cisco.com

Version of September 28, 2009

The paper entitled *Assessment of Student Learning in Science Simulations and Games* by Edys Quellmalz Quellmalz Quellmalz, Quellmalz, Timms, & Schneider (2009) presents a broad and valuable overview of the emerging landscape for simulation, games and assessments in the particular context of science assessment. The paper is informative and pulls together many pieces of a divergent literature. I especially appreciated their use of the language of Evidence Centered Design (Mislevy, Steinberg, & Almond, 2003) as I think it is the most promising framework for moving the instructional, assessment, and gaming communities forward together (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008).

I will respond to the questions which the committee has put to me by providing focus on some areas necessarily given limited treatment in the Edys Quellmalz Quellmalz Quellmalz et. al. (2009) paper.

Is the idea of using simulation or game-based assessment of science learning worth pursuing (i.e., does it add sufficient value over current practices to warrant investment)?

Assumptions

The degree to which this emerging genre is worth pursuing will depend on one's understanding of 'educational assessment', one's understanding of quality of the current offerings in educational assessment and the incremental value of simulation of game-based assessment. The cost of not pursuing should be considered as well.

The definition of 'educational assessment' is important because it often comes with highly political and evaluative assumptions (Smith & Fey, 2000). For example, the ground-breaking book, *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001) starts with the sentence "Educational assessment seeks to determine how well students are learning and is an integral part of the quest for improved education" (p. 1). This is notable in it's effort to unify instruction and assessment from the outset (second half of the sentence) while also stating an understanding of assessment that is fundamentally evaluative. That is to say it is about "how well" rather than "how varied" or "in what ways". We would allow ourselves more latitude if we considered assessment a process of characterizing knowledge, skills, and other attributes relevant to educational goals, and leave the door open for those characterizations to be used for evaluative purposes or not, depending on the need. For example, a teacher may take precious time out of the day to observe the clothing and physical condition of high risk students, not to learn "how well they are learning", but possibly only to understand how "well they are living". Insofar as the general well being of students is a concern of educators and educational institutions, this and many other socially impactful use cases are noticeably omitted from most discussions of "educational assessment". We should keep in mind that there is a long tradition in the psychometric and educational community, along with political biases for funding, for reinforcing a large scale and evaluative model (Gipps, 1999).

Our understanding of the scope of education and what counts as curriculum is important too. If our understanding is that one goal of science education is to increase the number of future scientists, then assessment of "awe and wonder for science", "appreciation of science", "connectedness of science to other disciplines" and "likeableness of scientists" may all be reasonable conceptualizations to characterize students or groups of students. If we think education should be concerned with these personal and policy issues, then so too should educational assessment.

What do we have (mostly) now?

While some comprehensive reviews of the current landscape of educational assessment point out some positive effects of the current approaches to large scale testing systems and common practice (Hamilton, 2003), the preponderance of educational policy literature takes a more negative or tentative stand (Gipps, 1999; Hamilton, 2003). Some of the most relevant concerns for this discussion center on the negative effects of an external and evaluative force in the classroom ((Nichols & Berliner, 2008a, 2008b; Shaul & Ganson, 2005; Sloane & Kelly, 2003; Smith & Fey, 2000), errors or negative effects in cut-score setting or other decision making and communication (Linn, 1998; Shepard, 1997; Sheppard, 2002) and informational concerns regarding the limitations of commonly used fixed response formats as compared to the complex and contextualized nature of classroom and out-of classroom life (Baxter, Shavelson, Goldman, & Pine, 1992; Hattie et al., 1999). An important interplay occurs between the assessment format and the evaluative function/curricular mis-alignment in the classroom. When we consider the broad range of graphic, auditory & written & kinesthetic representations that occur in the intellectually economy of the classroom, these are typically very different from the representational organization of the materials with which individuals interact in assessment. That is to say, the "outside" fixed-response format of many assessments requires an unnatural cessation of normal instructional experience to create assessment events. This leads to a teach/teach/teach/test/test/test pattern rather than a natural flow of instructional activity that includes instructional presentation, exploration and feedback.

Incremental Value

Desired areas of incremental value for new assessments can be gleaned from a list of desired features of authentic assessment (Darling-Hammond & Snyder, 2000):

> Assessments sample the actual knowledge, skills, and dispositions …rather than relying on more remote proxies. (2) Assessments require the integration of

multiple kinds of knowledge and skill as they are used in practice. (3) Multiple sources of evidence are collected over time and in diverse contexts. 4) Assessment evidence is evaluated by individuals with relevant expertise against criteria that matter for performance in the field.

Actual Knowledge Skills and Dispositions

As described by Edys Quellmalz Quellmalz Quellmalz et. al, appropriately created simulations have the potential to create rich micro-worlds in which tasks are naturally embedded and have clear evidentiary relationship to the kinds of activity that occur in the educational context, and the kind of activities we wish to make inferences about beyond the specific setting. This means, for example, collecting data, graphing the data, and identifying explanations or writing a narrative about the results. This is called the presentation process in the ECD assessment model and has the potential to become increasingly "real" and sophisticated as technologies evolve. (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004; Mislevy, Steinberg, & Almond, 2002)

This is especially important as scientific artifacts, data, representations, and modes of discourse become increasingly digital. As the world continues to evolve in this way, the natural representations for instruction and assessment would converge, making question-based evidence gathering increasingly more limited as it becomes replaced by electronically measured tasks. Aside from classroom tasks, it is important to recognize that the typical student will spend increasingly large amounts of time communicating with others and working by themselves using digital systems and representations. For the digital natives that are 15 years old or younger, they have never been alive without the World Wide Web.

Multiple Sources of Integrative Data Collected Over time

Behrens, et al. (2008) argued that games and assessments can be understood to have a common delivery structure. Using the ECD language the process is described as “select an activity”, “present the activity”, “capture and score the output”, and “summarize the score with previous scoring information”. “Repeat”. In an interesting game, the activity selection is optimized to maintain motivation and interest. In an assessment, the selection activity is typically tuned to maximize an information function. In a tutoring environment, the selection can be considered as optimizing tasks for transfer in the Zone of Proximal Development. Regardless of the purpose, the user experiences a series of cycles of get a task, do it, get feedback (or not) and get another task.

This symmetry provides the basis for the possibility that student could be involved in ongoing games that are closely aligned with the instructional goals of a course, but compelling enough to be engaging and motivating. By properly combining interesting narrative structure (as a good teacher would) with appropriate activity and interaction, the student may progress with higher self motivation.   Specific skill deficits could be addressed by intelligent tutor agents (human or software) that inject additional diagnostic tasks as part of the game play when needed.

What if we don’t pursue games and simulations?

In judging the next steps for policy and implementation, we need to consider not only the current state of assessment and the future state of instruction, but the future state of general activity of population. Our work becoming increasingly accomplished by digital means, and these skills will be increasing in importance. Likewise the representations of our daily interactions are increasingly being provided by digital means and our tasks are becoming increasingly geographically distributed and asynchronous. These are not fringe side effects of education, these are core changes in our society and the nature of our work. There is a cultural and historical imperative to consider careful and deliberate moves in the direction of simulation and game based assessment because we are moving with similar experience in many other areas of the life of our populace.

What are the most promising and/or immediate applications of simulation and game-based assessments in science learning?

As hinted at in the section above, the most promising applications are those that have a natural verisimilitude to authentic work and classroom activity. In the beginning, this may suggest assessment occur in small chunks organized around particular digital or partially -digital work tracks. For example, small complete multi-media objects may be created to capture computer-appropriate data for specific sets of tasks and sent to remote or local scoring and summary systems with the assessment affordances becoming increasingly digital as the classroom instruction becomes increasingly more digital as well.

As instructional interaction becomes increasingly based on digital artifacts, there is increased opportunity for what we in the Cisco Networking Academies have been calling a CAGI (pronounced “CAGEY”) architecture – a computing infrastructure that provides Curriculum, Assessment, & Gaming Integration. Historically, the curriculum team made “content” that was “loaded” into a media shell for students to navigate. Architecturally, the curriculum was always differentiated from the “true assessment system” because it did not save data while “tests” in the assessment system did. This occurred even though there were “quizzes” in the curriculum and our Packet Tracer simulation software (Frezzo, et. al, in press) was embedded throughout the curriculum with a complex ECD back-end for work product scoring and student directed feedback.

We started discussing the idea of transferring performance data from the chapter-embedded simulation activities to a business intelligence dashboard that would help instructors and students make sense of the large amount of performance data that is generated in such student interactions. Soon we recognized the fact that not only could the simulation send data to update the instructor or student, but the curriculum and the interactive objects in the curriculum could generate data as well. For example, sections of the curriculum itself could become assessment objects communicating motivational data regarding amount and quality of interaction with different aspects of the system prior to on-line chats, game play, or formal classes. Similarly, ongoing gameplay could be created in a series of activities that parallel (or pre-sage) class discussion and presentation. This is a work in progress.

The idea thereby develops to move from assessment as an intrusive burden to assessment as an ubiquitous unobtrusive aspect of the digital environment. If created appropriately, the motivation of the learners in the system and the breadth of experiences we afford them may be able to shift the experience to be more like those that occur in informal learning outside the classroom (Greenfield, 2009; Meltzoff, Kuhl, Movellan, & Sejnowski, 2009), than traditional assessments within the classroom.

It is important to note, that if such systems were to be created in on-line conditions, a number of additional affordances could be generated. For example, the game could be played as a multi-user on-line game and aspects of cooperation and competition may be able to be inserted. These elements could be arranged to match students into asynchronous groups or intact cohorts depending on the instructional goal. On-line games would also have the benefit of generating analyzable digital data that could become the basis for applying machine algorithms on performance logs or other work products. For example, (DeMark & Behrens, 2004), analyzed the router logs of engineering students using statistical natural language processing techniques to look for frequency patterns and programming command clusters across novice and expert groups.

These techniques allowed the identification of some patterns of performance that indicated the original scoring rules based on expert judgment were incorrect and the assumptions regarding the relationship between patterns of thought and observable outcomes had to be re-examined.

The implications of this digital and networking shift are important to understand. As instructional activities become increasingly digital, data to support the creation and validation of scoring rules and norm based feedback to instructors and students (e.g. "only .5% of successful students used that command, you might want to talk to the student about what she was thinking) will become increasingly available. As activities and data scale, there will be increased incremental value in having algorithms for rule induction and pattern recognition.

What are the barriers to implementing simulation or game-based assessments of science learning?

Barriers to move forward in this area will be along two major dimensions: Socio-cultural and scientific/technical. The first difficulty would be cost and co-operation. (Hamilton, 2003) noted that "the U.S. General Accounting Office (2003) estimated the cost to states of implementing NCLB using only multiple-choice tests at approximately $1 .9 billion, whereas the cost if states also include a small number of hand-scored open-response items such as essays would be about $5.3 billion". Moving forward, states and federal government may save money by sharing assessment and task designs and as well as data and reporting infrastructure. Typically such cooperation is haphazard. However, in these cases, increased data would provide more scalable cost-benefit ratios.

Second, detailed analysis of possible connection to unanticipated social side-effects would need to be examined. For example, a number of researchers have identified extreme gamers who exhibit a number of anti-social behaviors. Would future versions of "studying to the test" be correlated with negative social behaviors that can arise from pathological game use

(Charlton & Danforth, 2007; Morahan-Martin & Schumacher, 2000)? Likewise, we will have to monitor whether there are gender-based, or other sub-group related, side effects as well. For example, there has been a long history of gender differences in highly spatial tasks. This together with the gender differences in use of video games may increase differential assessment performance as well. (Greenfield, 2009) encouraged consideration of simulation based games ("video games") that occur in naturalistic contexts as part of the learning in informal environments (Bell, et. al, 2009). She cited the work of Rosser et al (2007) who reported high correlation between surgical performance in laparoscopic surgery with amount of experience with video games and video game performance. These types of interaction would have to be studied so that digital assessments with high stakes do not create differential impact. Likewise, the effect of different physical and mental limitations along with variation in culture and language would have to be understood as well.

Third, understandings and constraints of the political system would need to actively manage any divergence between the perception of seemingly more "valid" assessments (a construct that cannot be defined independent of use) and the understanding of the political community regarding its appropriate use. We would need to be careful not to repeat the mistakes we have made in the past regarding poor consequential validity with fixed-response exams, now applied to more sophisticated technologies.

Fourth a comprehensive change management approach would want to include the need for large scale teacher professional development. This could occur in a number of ways including instruction and practice related to how to use simulations and games for instructional purposes so there is a representational alignment when the assessment uses are later highlighted. Strategies regarding the position of instructors would have to be considered. For example, in our network simulator used in the Cisco Networking Academies, there is a comprehensive assessment authoring interface that is available to instructors, thereby empowering them both to use simulation based assessment as well as to customize or create their own.

Technical issues related to simulation based games and assessment will require advancement in all areas of assessement delivery: activity selection, presentation, evidence identification (task-scoring) and evidence accumulation (score aggregation). Key to all this work will be continued advancement on understanding the relationship between attributes of complex tasks and the evidence they provide and how those observations can be conceptually connected. As simulated micro-worlds increase in feature complexity, the burden for design clarity around tasks increases: it becomes easier to have complex things occurring for which there is no understanding or modeling. The "physical" modeling of the simulation will have to evolve in parallel with socio-motivational-cognitive modeling at the same time.

Despite the many advances in psychometrics over the last years, there are still no universal and straight forward methods for dealing with the complex of combinations, correlations and task pathways that may occur in a simulation based game. As noted above, continued research in the use of flexible statistical models and machine learning will complement the prospect of large amounts of data.

References

Baxter, G. P., Shavelson, R., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment.*JOURNAL OF EDUCATIONAL MEASUREMENT, 29*(1), 1-17.

Behrens, John T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2008). Structural, Functional, and Semiotic Symmetries in Simulation-Based Games and Assessments. In *Assessment of Problem Solving Using Simulations* (pp. 59-80). New York: Earlbaum.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to Evidence Centered Design and Lessons Learned From Its Application in a Global E-Learning Program. *International Journal of Testing*, *4*(4), 295-301. doi: Article.

Charlton, J. P., & Danforth, I. D. (2007). Distinguishing addiction and high engagement in the context of online game playing.*Computers in Human Behavior*, *23*(3), 1531-1548. doi: 10.1016/j.chb.2005.07.002.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, *16*(5-6), 523-545. doi: 10.1016/S0742-051X(00)00015-9.

DeMark, S. F., & Behrens, J. T. (2004). Using Statistical Natural Language Processing for Understanding Complex Responses to Free-Response Tasks.*International Journal of Testing*, *4*(4), 371-390. doi: 10.1207/s15327574ijt0404_4.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J.(in press). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors.*The Journal of Science Education and Technology.*

Gipps, C. (1999). Socio-cultural aspects of assessment. *REVIEW OF RESEARCH IN EDUCATION, 24 1999*, *24*, 355-392.

Greenfield, P. M. (2009). Technology and Informal Education: What Is Taught, What Is Learned. *Science*, *323*(5910), 69-71. doi: 10.1126/science.1167190.

Hamilton, L. (2003). Assessment as a Policy Tool.*Review of Research in Education*, *27*, 25-68.

Hattie, J., Jaeger, R., & Bond, L. (1999). Persistent methodological questions in educational testing. *REVIEW OF RESEARCH IN EDUCATION, 24 1999*, *24*, 393-446.

Linn, R. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting.*APPLIED MEASUREMENT IN EDUCATION*, *11*(1), 23-47.

Meltzoff, A. N., Kuhl, P. K., Movellan, J., & Sejnowski, T. J. (2009). Foundations for a New Science of Learning. *Science*, *325*(5938), 284-288. doi: 10.1126/science.1175626.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-67.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477-496. doi: 10.1191/0265532202lt241oa.

Morahan-Martin, J., & Schumacher, P. (2000). Incidence and correlates of pathological Internet use among college students. *Computers in Human Behavior*, *16*(1), 13-29. doi: 10.1016/S0747-5632(99)00049-7.

Nichols, S., & Berliner, D. (2008a). Why has high-stakes testing so easily slipped into contemporary American life?*PHI DELTA KAPPAN*, *89*(9), 672-676.

Nichols, S., & Berliner, D. (2008b). Testing the joy out of learning.*EDUCATIONAL LEADERSHIP*, *65*(6), 14-18.

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

Shaul, M. S., & Ganson, H. C. (2005). The No Child Left behind Act of 2001: The Federal Government's Role in Strengthening Accountability for Student Performance*Review of Research in Education*, *29*, 151-165.

Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools*, *34*(2), 85-97. doi: 10.1002/(SICI)1520-6807(199704)34:2<85::AID-PITS2>3.0.CO;2-R.

Sheppard, L. A. (2002). The Hazards of High-Stakes Testing. *Issues in Science & Technology, 19*(2), 53. doi: Article.

Sloane, F., & Kelly, A. (2003). Issues in high-stakes testing programs. *THEORY INTO PRACTICE*, *42*(1), 12-17.

Smith, M., & Fey, P. (2000). Validity and accountability in high-stakes testing. *JOURNAL OF TEACHER EDUCATION*, *51*(5), 334-344.