Recommendations for Technology and Innovation in Assessment

Edys S. Quellmalz, Michael J. Timms, Barbara C. Buckley

This paper elaborates and explains recommendations offered in the slide presentation. In addition, the paper provides references to publications and projects on which the recommendations are based.

Question 1: How can innovative technologies be deployed to create better assessments?

RECOMMENDATIONS

Our overarching recommendations are presented in the first slide:

- Break the mold! Transform; don't transition.
- Go beyond delivery, scoring, and reporting
- Focus new development on what is not currently well tested in paper formats, i.e., integrated knowledge, active processes
- Take advantage of capabilities of technology to represent domain systems and models
- Support use of "tools of the trade"
- Reform test form designs and timing
- Form collaboratives to develop collections of innovative tasks
- Create common core of state and classroom standards, specifications, task banks
- Create common platforms for authoring and administration

What is Tested

To gather evidence of student progress on rigorous standards, the new generation of technology-enabled assessments of student learning should "break the mold" of traditional testing methods. Early uses of technology in large-scale assessments tend to focus on economic savings and logistical efficiencies related to delivery, scoring and reporting (Quellmalz & Pellegrino, 2009). But the significant advantage offered by technology-enabled assessment is to support the measurement of "what" is tested, particularly integrated knowledge and challenging standards not measured well, or at all, in paper-based tests (Quellmalz & Haertel, 2004). Both the static modality of traditional tests and the constrained item formats limit measurement of the types of significant, recurring problems and goals called for in standards. Extended problem solving and inquiry within authentic, real-world tasks are seldom tested. Active, iterative problem solving of tasks with alternative approaches and solutions are not tapped. Sustained literacy tasks involving seeking, selecting, composing, revising, interpreting, presenting, and critiquing are not provided. Use of multiple sources and media are not possible. In science, traditional paper-based tests do not represent the causal, temporal, and dynamic interactions within systems in the natural world (Buckley, Gobert, Horwitz, & O'Dwyer, in press, 2009; Gobert & Buckley, 2000). In the designed world, engineering systems thinking and design

problems involving proposals for alternative designs, testing them, and evaluating tradeoffs are not typically well tested. Collaboration, a crucial 21st century skill, is not tested with real or virtual peers and experts.

The new generation of technology-enabled assessments can move past items testing decontextualized, discrete knowledge of simple facts and concepts. Innovative tasks can give greater emphasis to assessing understanding of the models and organizational structures and types of strategic reasoning within subject domains and their application to situations. In science, technology can organize innovative tasks to address grade appropriate models of systems in life, physical, and earth science. English language arts literacy tasks may be clustered within broad categories of narrative, persuasive, and informative discourse aims and generic discourse structures employed to achieve communication purposes. In mathematics, prototypical problem types can embed component skills.

Importantly, technology-enabled assessments allow design of innovative tasks in which students use technologies that are "tools of the trade" in the domain and that are routinely employed in postsecondary education and the work place. These tools support new levels of thinking and reasoning by broadening methods for finding and collecting information and data and for using tools to manipulate information and data during problem solving and interpretation. Information and communications technologies such as web browsers, word processors, editing, drawing, and multimedia programs support research, design, composition, and communication processes. These same tools can expand the cognitive skills that can be assessed, including planning, drafting, composing, and revision. In science, technology, engineering and mathematics (STEM), tools of the trade would include simulations, models, and visualizations, and tools for data collection, representation, and analysis. Innovative assessment tasks could elicit evidence of students' problem solving, inquiry, and decision making processes, and multiple appropriate solutions, as well as proficiencies with the tools.

Slides 3-8 describe the increasing use of innovative, technology-based tasks in major large-scale national and international assessments and their potential in a new generation of formative and summative tests. Online testing now occurs in numerous international, national, and state assessment programs. The 2009 Programme for International Student Assessment (PISA) included electronic texts to test reading, and in 2006 PISA conducted a pilot of computer-based assessment in science. The National Assessment of Educational Progress (NAEP) studied online versions of mathematics and writing tests in preparation for transitioning NAEP to electronic administrations in the near future (Sandene et al., 2005). Currently, over 27 states have operational or pilot versions of online tests for their statewide or end-of-course exams. This includes Oregon, which pioneered online statewide assessment, North Carolina, Utah, Idaho, Kansas, Wyoming, and Maryland. The 2011 NAEP writing assessment will require use of word processing and editing tools to compose essays. In professional testing, architecture examinees use computer assisted design programs (CAD) as part of their licensure assessment. The 2012 NAEP Technological Literacy Framework lays out examples of assessment targets, task scenarios and illustrative tasks that will guide the development of innovative tasks to be computer

delivered that relate to Technology and Society, Design and Systems, and Information and Communication Technology (ICT) (naeptech2012.org).

Slides 5-11 propose how the capabilities of technology can support design of innovative formative and summative assessments. Examples from the NSF-funded Calipers II project within WestEd's SimScientists program illustrate formative uses of technology to provide immediate, individualized feedback and coaching (Quellmalz, Buckley, & Timms, 2009). Examples of the simulation-based tasks also illustrate ways that cyber literacy and mathematics cyberlearning can be assessed in the context of science investigations.

How Testing is Conducted

Technology can permit administration of alternative test designs. Tests no longer need to be given at one point in time, but can be administered during the school year as students complete units of study. Student performances during extended projects can be sampled from component tasks during research, problem solving, and communication.

Technology enables standards-based curriculum embedded formative assessments, end of unit benchmark assessments, that can supplement, even replace, large-scale summative assessments. Common standards-based specifications for designing assessment tasks can connect classroom and state level assessments. To be formative, assessments must be administered *during instruction* and *used* by teachers and students to interpret progress and make adjustments (Black and Wiliam, 1998). Interim testlets administered periodically, but not used in ongoing instruction are not formative and should not be confused with formative purposes.

The new generation of student assessments will benefit from collaborative efforts that share expertise and costs (Quellmalz & Moody, 2004). State assessment systems need to be balanced by articulating the standards and assessment tasks and items used at multiple levels of the system. Development of collections of innovative tasks can support sharing within and between states and reduce costs, as well creation of common platforms for authoring and administering assessments. Summative test designs should consider use of multiple forms and matrix sampling. Assessment should become bi-directional, using evidence from classroom unit benchmark assessments aggregated up the state data system, and state-based tasks embedded within classroom assessments. These recommendations are addressed in more depth in Question 3.

Question 2. We envision the need for a technology platform for assessment development, administration, scoring, and reporting that increases the quality and cost-effectiveness of the assessments. Describe your recommendations for the functionality such a platform could and should offer.

Question 4. For technology platforms, address cost issues.

RECOMMENDATIONS

To maximize access and utility, any technology platform for developing, administering, scoring, and reporting results from innovative assessments should be Web based. It should allow access to the administration, scoring, and reporting aspects of the system from all standard web-browsers (with appropriate plug-ins such as Flash) and should not require the installation of any additional software on school computers. This will avoid many complex issues in setting up computers in schools to be able to access the assessment system.

As far as possible, the scoring in the innovative assessments should be computer-based, regardless of the item format. As Quellmalz and Pellegrino (2009) have noted, "A transformative advance in large-scale testing programs is the machine scoring of essays and constructed responses, including testing programs for the military, industry training, higher education admissions, and statewide K-12 achievement. Computerized scoring of free-responses uses complex statistical methods and techniques such as Latent Semantic Analysis (LSA) (Landauer, Laham & Foltz, 2003). Pearson is in its second year of using Knowledge Analysis Technologies, based on LSA techniques, to pilot the automated scoring of 46,000 brief constructed responses for the Maryland School Assessment (MSA) science test. The Educational Testing Service (ETS) has developed E-rater for scoring essays and C-rater for scoring constructed responses and has deployed them in a variety of high stakes testing programs such as the GMAT.

Klein (2008) recently reviewed the literature on automated scoring methods and presented results from a study comparing hand and machine scoring of college-level, open-ended items of the type found on the Collegiate Learning Assessment. Findings across studies using a variety of machine scoring methods consistently show comparability of human and machine scoring at levels sufficient to warrant using computerized scoring alone, or as an augmentation to human scoring. " (Quellmalz & Pellegrino, 2009).

Given the expansion of the types of knowledge and skills that will be addressed in complex innovative assessments, it will be necessary to accept a wider range of forms of evidence of achievement generated from student responses to complex tasks in the innovative assessments. This wider range of types of evidence will require the adoption of new methods of processing and evaluation of resulting data. The current psychometric methods applied in educational testing are not sufficient for this purpose, and the field must look to methods from other fields that handle more complex data such as intelligent tutoring systems.

By making the scoring automated, it allows the reporting to students and teachers to be instant, thereby enabling formative assessment. True formative assessment happens in the classroom, by the teacher using the results of the assessment to inform her decisions about future instruction for individual students, groups of students, or the class as a whole. Slide 19 provides an example from the NSF Calipers II SimScientists project of an embedded assessment report generated by the simulation-based science assessment. The report classifies students into groups based on their responses during the simulation to content

and inquiry tasks and items. The report indicates students that need help, are making progress, or are on track. The teacher can generate student, group or class summaries. Slide 20 displays summary class results of the unit benchmark simulation-based assessment, with students placed into the four profiency levels currently reported on state tests.

(4) For the technology ``platform'' vision you have proposed, provide estimates of the associated development and ongoing maintenance costs, including your calculations and assumptions behind them.

Additional Costs

The ongoing work on the development and study of the innovative assessments and the technology platforms that support them is not at a stage where it is possible to give accurate costs for scaling up such systems. As is typical in advanced technologies, the costs of the initial systems will be high and because the assessments are more complex, the costs of developing them are also high. It is also known that, for large-scale administration, there are increased site administration costs due to the need for more skilled personnel than the typical exam proctors.

Reduction of Additional Costs

Given that start-up costs will be high, it would be extremely beneficial for groups of states to form collaboratives to develop innovative assessment tasks and items and the technologies needed to support them. Costs of innovative items can be controlled by creating templates and specificaton shells for their design to allow for rapid prototyping and testing and by creating components of the assessments that can be reused across multiple items. In addition, given that complex innovative assessments will be more expensive, states will need to choose which topics they are best suited to and develop them for those in which there is a definite added value to existing assessment item types. This might involve using matrix sampling of the population too, rather than having to administer them to every student.

Cost Savings

Once assessments and supporting systems are in place, there will be cost savings compared to the current assessment programs. Savings will result from there being no cost for printing and shipping of paper-based assessments, no shipping and scanning of 'bubble sheets', and no human scoring sessions, given that scoring has been fully automated. In electronic environments, it is also easier to add accommodations like large print or read-aloud (text to speech) and, once the tools are in place to provide these, the ongoing costs to do so are minimal. In addition, the results of accountability assessments could be sent electronically to students and parents, thereby reducing mailing costs.

Question 3. How would you create this technology platform for summative assessments such that it could be easily adapted to support practitioners and professionals in the development, administration, and/or scoring of high quality interim assessments?
Question 4. What are cost considerations?

RECOMMENDATIONS

Slides 16-22 present our recommendations for developing balanced state assessment systems. We first emphasize the important distinction between interim assessments and formative assessments. Interim assessments typically sample from the state test and are given periodically, but are not scheduled to coincide with instructional units. Formative assessments target the knowledge and skills in a particular unit. They are designed to be used during instruction to gauge student progress and adjust instruction accordingly. Many published products of testlets are not used formatively by teachers during instruction and may be very limited by the formats of items.

In contrast, the formative assessments developed in the SimScientists projects include not only online assessments embedded in instruction, but also progress reports to the teacher and students, and follow up off-line classroom reflection activities. The online assessments provide students with immediate feedback and multiple levels of coaching based on their actions and answers. The progress report identifies the concepts for which student understanding is on track, in development, or needing help. Based on the progress report, the teacher assigns students to teams. Students who need help in a key concept are assigned to a team that applies that key concept in a new context. Similarly, students whose understanding is under development are provided with a task that will facilitate that development. Students who have mastered the content are given a task that asks them to stretch and articulate the more difficult concepts of the unit. Students engage in scientific discourse focused on observation and evidence. The different teams then come together in larger groups to integrate their understandings and present their evidence and conclusions to their fellow students. Thus, teachers are given the information they need to understand where their students are having difficulties in mastering the concepts and skills, and materials that enable them to assign tasks that will facilitate the development of student understanding.

The summative unit benchmark assessments are end-of-unit online assessments that assess student understanding with task types similar to those used in the embedded formative assessments, but presented in a new context. The key differences between the embedded formative assessments and the summative benchmark assessments are [1] the absence of feedback and coaching during the online assessment and [2] a proficiency report that characterizes student performance on key concepts and skills in NCLB proficiency categories. Tasks and items in the benchmark assessments also tend to be more integrated than those in the embedded formative assessments, because we are not so constrained by diagnosing and providing feedback and coaching for weaker performances. (SimScientists project descriptions, publications, and examples may be viewed at http://simscientists.org)

Our recommendations for balanced, multilevel state assessment systems are drawn from a National Academy paper, "Developing Multilevel State Science Assessment Systems" and ongoing research and development projects funded by the U.S. Department of Education Institute of Education Sciences and OESE: Multilevel Assessments of Science Standards (MASS) and Integrating Science Simulations into Balanced State Science Assessment Systems. Our work in science is studying the use of design templates, specification shells, storyboards, and re-usable components for rapid and cost-effective development. In the Enhanced Assessment Grant, a Design Panel of six states (CT, MA, NC, NV, UT) led by Nevada is studying the feasibility, utility, and technical quality of simulation-based benchmark assessments for inclusion in a state's report on achievement of science standards (Qullmalz & Silberglitt, 2009). That project and the MASS project are also studying the effects of the simulation-based formative curriculum-embedded assessments on subsequent performance on the unit benchmark assessment and district and state science tests. Findings from these projects will inform questions about the potential role and utility of innovative assessments in state science assessment systems.

We consider a key strategy for linking classroom formative and state tests the creation and use of common task design specifications for core tasks at state and classroom levels. We also propose that state collaboratives develop and share a common core collection of secure and public tasks to link and support assessments across the levels. Finally, we recommend design and study of a variety of models for constructing assessment systems that could, for example, take advantage of unit benchmark assessments with established technical quality by aggregating them into state achievement data, or where secure state developed tasks could be embedded in unit benchmark assessments. All of these efforts can take advantage of technology to change in fundamental ways the what, how, when, and where of testing.

REFERENCES

Buckley, B. C., Gobert, J., Horwitz, P., & O'Dwyer, L. (in press, 2009). Looking inside the black box: Assessing model-based learning and inquiry in BioLogica. *International Journal of Learning Technologies*.

Black, P. & Wiliam, D. (1998). Inside the black box: *Raising standards through classroom assessment.* London, UK: King's College.

Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. *International Journal of Science Education, 22*(9), 891-894.

Klein, S. (2008) in Probability and Statistics: Essays in Honor of David A. Freedman, D. Nolan, T. Speed, Eds. (Institute of Mathematical Statistics, Beachwood, OH, 2008), vol. 2,pp. 76–89.

Landauer, T.D, Laham, D.,Foltz, P. (2003). Assessment in Education. 10, 295 (2003).

National Science Foundation. (2009). *Cyberlearning.*

Quellmalz, E.S., Timms, M.J., & Schneider, S.A. (2009). *Assessment of Student Learning in*

Science Simulations and Games. Paper commissioned by the National Academy of Science.

Quellmalz, E.S., Timms, M.J., & Buckley, B.C. (in press). The promise of simulation-based Science Assessment: The Calipers Project. *International Journal of Learning Technologies.*

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

Quellmalz, E. S., Buckley, B. C., & Timms, M. J. (2009). *Using Simulations to Support Powerful Formative Assessments of Complex Science Learning*. Paper presented at the NARST.

Quellmalz, E. S., & Haertel, G. D. (2008). Assessing new literacies in science and mathematics. In D. J. Leu, Jr., J. Coiro, M. Knowbel, & C. Lankshear (Eds.) *Handbook of research on new literacies*. Mahwah, NJ: Erlbaum.

Quellmalz, E. S. & Moody, Mark. (2004). *Models for multi-level state science assessment systems.* Report commissioned by the National Research Council Committee on Test Design for K-12 Science Achievement.

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.

Sandene, B. et al., "Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project" (NCES 2005-457, U.S. Department of Education National Center for Educational Statistics,U.S. Government Printing Office. Washington, DC, 2005).