Preparing for the Future: What Educational Assessment Must Do
A Summary Paper

Randy Elliot Bennett

Educational Testing Service

Princeton, NJ 08541

rbennett@ets.org

Paper presented at the annual meeting of the International Association for Educational Assessment, Astana, Kazakhstan, September 2012.

There is little question that education is changing. The means by which individuals learn are shifting from traditional ones to electronic media. Witness the rise of educational games, as well as the attention being given to those games by the academic community (e.g., Gee & Hayes, 2011; Shaffer & Gee, 2006). Simultaneously, what individuals must learn is evolving due to an exponential accumulation of knowledge and of technology to access, share, and exploit that knowledge. In the US, the re-conceptualization of school competency in the form of the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) signals one attempt to respond to that change. Finally, how education is organized, offered, and administered is undergoing transformation. The possibility of assembling one's post-secondary education from free Internet course offerings, with achievement documented through certification "badges," appears to be rapidly coming to reality (Young, 2012).

With such potentially seismic changes in education must also come changes in educational assessment (Bennett, 2002). This paper summarizes 13 claims about what educational assessment must do if it is to remain relevant.

1. Provide Meaningful Information

It should be obvious that in order to make sensible decisions about the effectiveness of education systems and the preparedness of populations, policy makers need meaningful information. Similarly, teachers and students need meaningful information if they are to effectively plan and adjust instruction. To be relevant, future educational assessment systems will need to provide trustworthy and actionable summative information for policy makers and formative information for teachers and students.

2. Satisfy Multiple Purposes

The previous claim indicated that educational assessment must provide meaningful information for summative and formative purposes. That claim is somewhat oversimplified because the demand for meaningful information centers upon *multiple* summative and *multiple* formative purposes. Education officials demand information to assist in evaluating students for promotion and graduation; schools (and school staff) for rewards and sanctions; and intervention programs for continuation and expansion. Educators also demand more fine-grained information for deciding what to teach when to whom; for helping teachers refine their instructional practice; and for improving educational programs.

This array of purposes cannot possibly be satisfied with a single test. Multiple purposes might best be served by different, related assessments designed to work in synergistic ways—i.e., through modular systems of assessment. The modular systems approach is taken by the Smarter Balanced (Smarter Balanced Assessment Consortium, 2010) and Partnership for Assessment of Readiness for College and Careers (2010) assessment consortia, as well as in such research initiatives as Cognitively Based Assessment of, for, and as Learning (Bennett, 2010; Bennett & Gitomer, 2009).

3. Use Modern Conceptions of Competency as a Design Basis

Across competency domains, the knowledge, processes, strategies, and habits of mind that characterize communities of practice differ fundamentally. At the same time, there are competencies that appear to be more general (Gordon, 2007). Our knowledge about the nature of these general as well as domain-based proficiencies is constantly evolving. In addition, the proficiencies our society considers to be important are evolving. The implication is that assessment design must be firmly grounded in up-to-date conceptions of what it means to be a proficient performer within valued domains, as well as in those competencies that have more general applicability (including socio-emotional ones). Concentrating on only domain-based competencies or only focusing on general ones will not suffice (Perkins & Salomon, 1989).

4. Align Test and Task Designs, Scoring, and Interpretation with Those Modern Conceptions

Grounding design in modern conceptions of competency means, at the least, developing competency models that propose what elements make for proficiency in a domain (and across domains), how those elements work together to facilitate skilled performance, and how they might be ordered as learning progressions for purposes of instruction. Second, it means extracting from research a set of principles for good teaching and learning practice to guide assessment design. Finally, it means developing an assessment design, the tasks composing it, and mechanisms for the scoring and interpretation of examinee performance that are logically linked to the competency model, learning progressions, and/or principles for good teaching and learning practice. That linkage should be documented in a detailed design document that becomes part of the interpretive argument for the test (Kane, 2006).

An important implication of aligning with modern conceptions of competency is that educational assessment will need to go well beyond traditional item formats. Modern conceptions recognize the importance of posing reasonably realistic problems that require students to exercise control over multiple competencies simultaneously. Such conceptions will make mandatory the use of more complex tasks, including simulations and other extended constructed-response formats.

5. Adopt Modern Methods for Designing and Interpreting Complex Assessments

To align design, scoring, and interpretation to modern conceptions of competency, we will need to adopt modern methods. Methods such as Evidence-Centered Design (ECD) (Mislevy, Almond, & Lukas, 2003) offer well-founded inferential structures and mechanisms to aid in the creation of assessments and in making sense of the results. Frameworks like ECD offer: (1) a way of reasoning about assessment design, (2) a way of reasoning about examinee performance, (3) a data framework of reusable assessment components, and (4) a flexible model for test delivery.

6. Account for Context

A student's performance on an assessment--i.e., the responses the student provides and the score the student achieves--is a fact. *Why* the student performed that way is an interpretation. For many decision-making purposes, that interpretation needs to be informed by an understanding of the context in which the student lives, learns, was taught, and was assessed.

This need is particularly acute for large-scale tests for which decisions typically center upon comparing individuals or institutions to one another, or to the same competency standard, so as to facilitate a particular decision (e.g., graduation, school accountability, postsecondary admissions). Because of the need to present all students with the same tasks administered under similar conditions, those tests will be far more distant in design, content, and format from the instruction students actually encounter. In this sense, such tests are "out of context." Context is provided by supplementary information such as, for college and graduate admissions, grade-point-average, transcripts, letters of recommendation, and personal statements.

Embedding assessment directly into the learning context should make assessment information more actionable for formative purposes. For a variety of reasons, this in-context performance might not be useful for purposes beyond the classroom or learning environment generating the data. The large number and wide diversity of such learning environments may make aggregated results meaningless. In addition, attaching significant consequences to activity in environments built to facilitate learning may unintentionally undermine both the utility of the formative feedback and achievement itself.

7. Design for Fairness and Accessibility

Among US social values is equal opportunity for individuals, as well as for traditionally underserved groups. In standardized testing, fairness for individuals was a motivating concern going back to the ancient Chinese Imperial civil service examinations. In modern measurement, concern for fairness for traditionally underserved groups dates at least to the 1930s, when Carl Brigham (1930) recognized "that tests in the vernacular must be used only with individuals having equal opportunities to acquire the vernacular of the test" (p. 165), if interpretation and use was to be meaningful.

Concern for fairness will continue regardless of the form that future educational assessments take. That concern will not be restricted to consequential tests but extend to formative assessment as well. Formative assessments entail a two-part validity argument: (1) that the formative instrument or process produce meaningful inferences about what students know and can do, leading to sensible instructional adjustments and (2) that these instructional adjustments consequently cause improved achievement (Bennett, 2011). Fairness would seem to require that this argument hold equally well across important population groups--that is, a formative assessment instrument or process should provide similarly meaningful inferences about student competency, suggest similarly sensible instructional adjustments, and lead to similar levels of instructional improvement.

8. Design for Positive Impact

It is generally acknowledged that, for consequential assessments, test design and use can have a profound impact--sometimes intended, sometimes not--on individuals and institutions (Koretz & Hamilton, 2006). Examples of impact may be on the behavior of teachers and students, or on the behavior of organizations (e.g., schools). *No Child Left Behind* was premised on intended positive impact—i.e., focusing educators in underachieving schools on the need to improve and, in particular, on improvement for underserved student groups.

Test design and use can also have unintended effects. In the case of *No Child Left Behind*, those effects are commonly asserted to include large amounts of instructional time spent "teaching to the test."

The reasoning behind the *Race to the Top Assessment Program* appears to be that, if low quality standards and narrow assessments can have negative effects, then high quality standards and assessments ought to be able to have positive impact (US Department of Education, 2010). Using principles and results from learning sciences research, summative assessments can be designed to model good teaching and learning practice by, among other things, giving students something substantive and reasonably realistic with which to reason, read, write, or do mathematics or science; routinely including tools and representations similar to ones proficient performers employ in their domain practice; and structuring tests so that they demonstrate to teachers how complex performances might be scaffolded.

9. Design for Engagement

Assessment results are more likely to be meaningful if students give maximum effort. Electronic game designers seem to have found ways to get students to give that effort.

Why not simply embed assessment into a game, thereby creating an engaging assessment? For formative purposes, that strategy might work to the extent that the game was designed to exercise relevant competencies and game play can be used to generate meaningful information for adjusting instruction. For summative purposes, game performance might offer useful information *if,* among other things, everyone plays the same game, or a common framework can be devised for meaningfully aggregating information across students playing different games intended to measure the same competencies. That latter model is employed in the *Advanced Placement Studio Art* assessment, for which students undertake different projects, all of which are graded according to the same criteria (Myford &, Mislevy, 1995).

10. Incorporate Information from Multiple Sources

All assessment methods-- tests, interviews, observations, work samples, games, simulations—*sample* behavior. Further, each method is subject to its own particular limitations, or method variance. In combination, these facts argue for the use of multiple methods. Multiple sources are commonly used for such consequential decisions as postsecondary admissions, where consideration is given not only to tests scores, but to grade-point-average, interviews, personal statements, and letters of recommendation.

To the extent practicable, this claim also would suggest using multiple sources of evidence for formative decision making. Rather than adjusting instruction on the basis of a single interaction or observation, the teacher would be wise to regard the inference prompted by that initial observation as a "formative hypothesis" (Bennett, 2010), to be confirmed or refuted through further data gathering.

11. Respect Privacy

In a technology-based learning environment, assessment information can be gathered ubiquitously and surreptitiously. Some commentators have suggested that this capability will lead to the “end of testing” (Tucker, 2012). That is, there will be no reason to have stand-alone assessments because the information needed for classroom, as well as accountability purposes, will come from learning and instruction.

Whereas this idea may seem attractive, individuals should know when they are being assessed and for what purposes. Second, having every learning (and teaching) action recorded for consequential purposes could potentially stifle experimentation in learning and teaching, including the productive making of mistakes (Kapur, 2010).

A compromise might be similar to the approach used in many sports. In baseball, the consequential assessment that counts toward player statistics and team standing occurs only during the official competition. Spring training, before-game practice, in-between inning practice, and in between-game practice are reserved for learning. We might consider doing the same for assessment embedded in learning environments—i.e., use separately identified periods for consequential assessment vs. learning (or practice).

12. Gather and Share Validity Evidence

However innovative, authentic, or engaging they may be, future assessments will need to provide evidence to support the inferences from, and uses of, assessment results. Legitimacy is granted to a consequential assessment by a user community and the scientific community connected to it. Among other things, that legitimacy rests upon the assessment program providing honest evaluation of the meaning of assessment results and the impact of the assessment on individuals and institutions; reasonable transparency in how scores are generated; and mechanisms for continuously using validity results to improve the assessment program.

13. Use Technology to Achieve Substantive Goals

The final claim is that future assessments will need to use technology to do what can't be done as well with traditional tests. Among those uses will be to measure existing competencies more effectively (and efficiently). A second use will be to measure new competencies, including features of the examinee's problem-solving process (Bennett, Persky, Weiss, & Jenkins, 2010). Third, technology might be deployed to have positive impact on teaching and learning practice.

Conclusion

Education, and the world for which it is preparing students, is changing quickly. Educational assessment will need to keep pace if it is to remain relevant. This paper offered a set of claims for how educational assessment might achieve that critical goal.

References

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning,and Assessment, 1(1)*. Available: http://escholarship.bc.edu/jtla/vol1/1/

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70-91.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice 18*, 5-25.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.*), Educational assessment in the 21st century* (pp. 43-61). New York, NY: Springer.

Bennett, R.E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment, 8(8)*. Retrieved from http://escholarship.bc.edu/jtla/vol8/8

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37*, 158-165.

Gee, J. P., & Hayes, E. R. (2011). *Language and learning in the digital age*. Milton Park, Abingdon, England: Routledge.

Gordon, E. W. (2007). Intellective competence: The universal currency in technologically advanced societies. In E.W. Gordon & B. R. Bridglall (Eds.), *Affirmative development: Cultivating academic ability*. Lanham, MD: Rowan & Littlefield.

Kane, M. (2006). Validation. In R.Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport , CT: Greenwood Publishing.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R.Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport , CT: Greenwood Publishing.

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38(6),* 523-550.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to Evidence-Centered Design* (Research Report 03-16). Princeton, NJ: Educational Testing Service.

Myford, C. E., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (MS-94-05). Princeton, NJ: Educational Testing Service.

National Governors Association Center for Best Practices & Council for Chief State School Officers. (2010). *Common Core State Standards.* Washington, DC: Author.

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the*

Top Comprehensive Assessment Systems Competition. Retrieved August 9, 2012 from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context bound? *Educational Researcher, 18(1)*, 16-25.

SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program application for new grants: Comprehensive assessment systems CFDA Number: 84.395B*. Retrieved from: http://www.k12.wa.us/SMARTER/RTTTApplication.aspx

Shaffer, D. W., & Gee, J. P. (2006). *How computer games help children learn*. Houdsmills, England: Palgrave MacMillan.

Tucker, B. (2012, May/June). Grand test auto: The end of testing. *Washington Monthly*. Retrieved from http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php

US Department of Education. (2010). *Race to the Top Assessment Program: Application for new grants*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop-assessment/resources.html

Young, J. (2012, January 8). 'Badges' earned online pose challenge to traditional college diplomas. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Badges-Earned-Online-Pose/130241/