The Gordon Commission
on the Future of Assessment in Education

Preparing for the Future: What Educational Assessment Must Do

Randy Elliot Bennett
Educational Testing Service

There is little question that education is changing, seemingly quickly and in some cases dramatically. The mechanisms through which individuals learn are shifting from paper-based ones to electronic media. Witness the rise of educational games, available on the personal computer, tablet, and mobile phone, as well as the attention being given to those games by the academic community (e.g., Gee & Hayes, 2011; Shaffer & Gee, 2006). Simultaneously, the nature of what individuals must learn is evolving, in good part 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 [NGO & CCSSO], 2010) signals one attempt to respond to that change. Finally, how education is organized, offered, and administered is undergoing transformation, most apparently--but not only--in higher education. 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 potentially seismic changes in the mechanisms, nature, and organization of education must also come changes in educational assessment (Bennett, 2002). Otherwise, education and assessment will work against one another in ever increasing ways. This paper offers a set of 13 claims about what educational assessment must do if it is to remain relevant but, even more so, if it is to actively and effectively contribute to individual and institutional achievement. The claims are that educational assessment must:

1. Provide meaningful information
2. Satisfy multiple purposes
3. Use modern conceptions of competency as a design basis
4. Align test and task designs, scoring and interpretation with those modern conceptions
5. Adopt modern methods for designing and interpreting complex assessments;
6. Account for context
7. Design for fairness and accessibility
8. Design for positive impact
9. Design for engagement
10. Incorporate information from multiple sources

11. Respect privacy
12. Gather and share validity evidence
13. Use technology to achieve substantive goals

Each of these claims is discussed in turn.

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. The implication of this claim is that, to be relevant, future educational assessment systems will need to provide trustworthy and actionable summative information for policy makers (including local administrators) as well as formative information for teachers and students.

For both types of assessment, the provision of "meaningful information" implies results that faithfully reflect the state of educational achievement for an individual or a group. That reflection will be at a finer grain size in the formative case and at a larger one for the summative case. "Faithfully" implies the existence of an evidentiary argument that supports the fidelity of that reflection (Mislevy, Almond, and Lukas, 2003). Ideally, that reflection should carry with it implications for action – whether micro-adjustments to learning or macro education-policy changes – which also should be supported by an evidentiary argument.

There is no indication that the need will subside for such information or for assessment mechanisms to provide that information. If anything, the need will increase because of the international competition enabled by a global economy and by the belief that a productive and educated workforce is central to maintaining (or increasing) one's standard of living in such an economy (Bennett and Gitomer, 2009). The rapid growth of international assessments is one indicator of this need for summative educational information. In 2000, 43 countries/economies participated in PISA, the *Programme for International Student Assessment*, sponsored by the Organizations for Economic Co-operation and Development (OECD, n.d. a). By 2012, 64 entities were being assessed (OECD, n.d. b). Interestingly, the proportional increase in participation was overwhelmingly due to an influx of *non*-OECD countries/economies, which

tend to be less economically developed than the Organisation's membership. In 2000, 14 of the 43 entrants were non-OECD members, whose participation was presumably motivated by the chain of reasoning stated above (i.e., an educated workforce leads to an improved standard of living). In 2012, 30 of the 64 participants were non-OECD members.

A similar case can be made with respect to the need for effective formative assessment. Interest in formative assessment has grown dramatically since publication of the 1998 position and review papers by Black and Wiliam (1998a, 1998b, 1998c). This interest is fueled by the belief that formative assessment actually does what proponents claim – i.e., causes large improvements in learning. Although the research findings and conceptual grounding underlying such claims have been over-stated at best (Bennett, 2011; Coffey, Hammer, Levin, & Grant, 2011; Kingston and Nash, 2011), to remain relevant the educational assessment community must do its best to produce tools and practices that do, in fact, enhance achievement. Educators expect it and students deserve it.

The question, then, is not whether summative and formative assessments will continue to be necessary but rather the form(s) they will take and the competencies they will measure, claims to which we will soon turn.

Satisfy Multiple Purposes

The previous claim indicated that educational assessment must provide meaningful information for summative and formative purposes. As stated, that claim is somewhat oversimplified because, in reality, 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.

It should be obvious that this array of purposes cannot possibly be satisfied with a single test because an assessment built for one purpose won't necessarily be suited to other purposes. Building an assessment to serve many purposes also is also unlikely to work because an assessment designed for multiple ends may prove optimal for none of its targeted goals. A

formative assessment used to generate summative information incidentally is likely to do a poor job at both purposes (for reasons to be discussed later). 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 the one 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 by such research initiatives as CBAL (Bennett, 2010; Bennett and Gitomer, 2009).

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 of this claim 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). Either a domain-based focus or a general focus alone will not suffice (Perkins and Salomon, 1989).

Unfortunately, the conceptions of competency that underlie many current tests, especially those used in primary and secondary school assessment programs, have their grounding in a behaviorist learning theory circa 1950 rather than in the modern learning sciences (Shepard, 1991). In general, those assessment programs do not directly measure knowledge construction, knowledge organization, knowledge schema, procedural fluency, the coordination and integration of competencies required for complex performance, and the problem-solving process, to name a few key constructs. Nor do those tests account for the qualitative orderings in competency development, or learning progressions, that are emerging from theory and research (Corcoran, Mosher, and Rogat, 2009; Daro, Mosher, and Corcoran, 2011; Educational Testing Service, 2012). Such progressions could potentially increase the relevance of test results for teachers and students.

One implication of this claim is that although content standards, such as the Common Core State Standards (NGA and CCSSO, 2010) help, those standards do not necessarily reflect findings from the learning sciences in ways that can effectively guide test design. A bridge from content standards to test design can be provided by competency models that identify the components required for successful performance within and across domains, and how those components might be organized; learning progressions describing hypothesized paths to competency development; and principles for good teaching and learning practice (Bennett, 2010). Describing the literature base underlying the models, progressions, and principles is a key to making the case for those entities as a credible design basis.

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

It is one thing to espouse grounding design in modern conceptions of competency. It is another thing to do it. Doing it 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, at least in the world of primary and secondary schools, is that educational assessment will need to go well beyond traditional item formats (Bennett and Ward, 1993; Pellegrino, Chudowsky, and Glaswer, 2001). Modern conceptions recognize the importance of posing reasonably realistic problems that call upon examinees to connect knowledge, processes, and strategies to conditions of use. Those conceptions also posit the importance of problems requiring students to exercise control over multiple competencies simultaneously, then deploying and integrating those competencies

in planful ways to achieve a desired result. Such conceptions will make mandatory the use of more complex tasks, including simulations and other extended constructed-response formats. That use, however, needs to be clearly motivated by the need to measure competencies that cannot be assessed through less labor-intensive means (or by some other important benefit).

Although modern conceptions of competency will make the use of complex tasks unavoidable, that use should not necessarily dominate. More elemental, discrete tasks are needed to decompose complex performance for formative purposes; i.e., to help teachers and students identify what subcompetencies might be responsible for failure on a complex task. For summative purposes, discrete items also can play a role by helping to reduce the impact of such unwanted task effects as lack of generalizability (Linn and Burton, 1994).

Finally, future scoring mechanisms, regardless of whether human or automated, will need to align with relevant domain processes. Ideally, more sophisticated scoring methods should bring with them the ability to recover the very knowledge structures, problem-solving processes, strategies, and habits of mind that tasks are designed to evoke. One might try to justify scoring responses through methods that don't attempt to account directly for the target competencies (e.g., machine learning, regression of human scores on nonaligned response features) but that justification would be a weak one.

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, and Lukas, 2003) and assessment engineering (Luecht, 2009) 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: a) a way of reasoning about assessment design, b) a way of reasoning about examinee performance, c) a data framework of reusable assessment components, and d) a flexible model for test delivery.

Reasoning about assessment design begins with specifying the claims to be made about individuals or institutions on the basis of assessment results. Those claims should derive directly from competency models and learning progressions. Specified next is the evidence needed to support those claims. Finally, the tasks required to elicit that evidence are described.

In assessment design, the reasoning chain is as follows: examinees whose competencies are consistent with a given claim will provide particular evidence in responding to the described tasks. Reasoning about examinee performance proceeds in the reverse direction. That is, when a given examinee offers evidence consistent with a claim in response to an aligned task, we can infer with some estimable level of uncertainty that the examinee meets the claim. As more task responses from that examinee are gathered to provide evidence about the claim, our belief in examinee standing with respect to the claim is updated and our level of uncertainty generally gets smaller.

Evidence is accumulated through a measurement model that generates a score, a qualitative characterization (e.g., a level in a learning progression, a diagnosis), or both. That measurement model also provides an estimate of the uncertainty associated with that score or characterization. The operational infrastructure in most large testing programs today can accommodate simple measurement models, generally models that array examinees along a single dimension. The operational infrastructure needs to be created for multidimensional models — i.e., models that extract evidence from an item for more than one dimension simultaneously.

Measurement models are only important, of course, if the purpose of assessment is to characterize student performance in some way that requires the notion of uncertainty. Inferences about some latent attribute of the student (e.g., that the student has achieved proficiency in some domain, or has a given standing with respect to some variable of interest), the likelihood that the student will perform acceptably in some other environment, or the likelihood that the student is a member of a particular diagnostic category all bring with them such uncertainty. In contrast, if the purpose of assessment is simply to judge a student's performance qua performance--as in an Olympic sporting event--without any attribution beyond describing the observed result, then no inference is needed, no uncertainty is implied, and no measurement model is required. That the student achieved a particular score or ranking in an event, and won a medal (or didn't) are facts. (See Messick, 1992, for discussion of these two situations in the context of performance assessment.)

A third benefit of modern design methods is the potential for a data framework of reusable assessment components. For example, task models can be created to specify the elements of a family of questions (e.g., competency model and learning progression claim, stimulus characteristics, stem characteristics, response format). Generalized rubrics then can be

created for scoring that family of questions (Bennett, Morley, & Quardt, 2000). Evidence model fragments that accumulate responses across some specified set of tasks can be generated. These task models, generalized rubrics, and evidence model fragments can, in principle, be stored in a data library. Creating a new assessment then proceeds by selecting data components that match the claims of interest.

A last design benefit is a flexible infrastructure delivery model. The four-process model consists of activity selection, presentation, response processing, and summary scoring (evidence accumulation). Creating the delivery infrastructure so that the four processes are separate allows for assembling new assessments, or changing old ones, in modular fashion. For example, the activity selection and presentation processes might be set to use members from the same task model in both a summative test and a diagnostic assessment but the response processing and summary scoring processes might be differently configured for those two use cases. For the summative case, response processing might extract a correct/incorrect judgment for each answer and accumulate across answers so as to estimate standing on a single dimension, whereas for the diagnostic assessment, aspects of the examinee's answer process might be judged and accumulated to produce a qualitative characterization.

Account for Context

A student's performance on an assessment – that is, the responses the student provides and the score the student achieves – is an indisputable fact. *Why* the student performed that way, and in particular, what that performance says about the student's competencies, is an interpretation. For many decision-making purposes, to be actionable, 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 (or types of tasks) administered under similar conditions, those tests, in contrast to classroom assessment, will be far more distant in design, content, and format from the instruction students actually encounter.

That distance is predicated upon the intention to measure competencies likely to manifest themselves across a variety of contexts, rather than in any particular one. In this sense, such tests are "out of context."

At present, our attempts to factor context more finely into the interpretation of large-scale test results take a variety of forms. In college and graduate admissions, for example, context is provided indirectly by grade-point-average and transcripts, and more directly by letters of recommendation and personal statements. These factors are combined clinically by admissions officials in decision making. For federal school accountability purposes, under *No Child Left Behind*, limited contextual data must be reported in addition to test-related information, including tabulations concerning "highly qualified teachers" and attendance and dropouts (State of New Jersey, Department of Education, n.d. b).

States may choose to compile additional information outside the requirements of *NCLB*. The complete New Jersey state "School Report Card" includes average class size, length of school day, instructional time, student/computer ratio, Internet connectivity, limited English proficiency rate, disability rate, student attendance rate, dropout rate, graduation rate, student suspensions and expulsions, student/faculty ratio, faculty attendance rate, faculty mobility rate, faculty and administrator credentials, National Board of Professional Teaching Standards certification, teacher salaries, and per pupil expenditures (State of New Jersey, Department of Education, n.d. a). Although New Jersey provides a wealth of information about the school-level context in which students are being educated, it offers no guidance about how to use that information for interpreting test results. Further, the state offers very little insight into the instructional context that characterizes any given classroom or into the home environment in which its students reside. How those factors should shade the interpretation of assessment results, and inform action, is left for teachers and parents to gauge for themselves.

Embedding assessment directly into the learning context – i.e., more closely integrating assessment with curriculum and instruction – should make assessment information more actionable for formative purposes. Such embedded assessments will be integral components of anytime/anywhere, online learning environments into which those assessments can be seamlessly fit. For a variety of reasons, this in-context performance might not be useful for purposes beyond the classroom or learning environment generating the data (e.g., for school accountability, college admissions, teacher evaluation). The large number and wide diversity of such learning

environments may not make aggregation meaningful. 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 (Black & Wiliam, 1998a). Last, the constant and potentially surreptitious surveillance of student behavior may pose privacy issues significant enough that some students opt out.

Design for Fairness and Accessibility

Among our country's social values is the idea of fairness in the form of equal opportunity for individuals, as well as for traditionally underserved groups. In standardized testing, fairness for individuals was a motivating concern from the earliest implementations of the practice, going back to the ancient Chinese civil service examinations (Miyazaki, 1976), which were instituted to ensure that jobs were awarded on merit rather than social class or family connections.

In the United States, concern for fairness did not originally extend to groups. In fact, several of the field's progenitors expressed racist views, perhaps most obviously in their interpretations of test results (e.g., Brigham, 1923) and most destructively in their failure to object to the use of their work to support racist and anti-immigration political agendas. Among the earliest statements of concern for group fairness from within the field was that of Carl Brigham (1930, p. 165) who, ironically, was a former eugenicist:

> For purposes of comparing individuals or groups, it is apparent that tests in the vernacular must be used only with individuals having equal opportunities to acquire the vernacular of the test. This requirement precludes the use of such tests in making comparative studies of individuals brought up in homes in which the vernacular of the test is not used, or in which two vernaculars are used. The last condition is frequently violated here in studies of children born in this country whose parents speak another tongue. It is important, as the effects of bilingualism are not entirely known.
>
> He went on:
>
> This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies – the writer's own – was without foundation. (p. 165)

Brigham's concern unfortunately did not take root for many years to come (with the notable exception of the *SAT*, which was instituted in the 1930s to *increase* access for economically diverse students to Harvard and other selective institutions [Bennett, 2005]). Among other things, tests were used well into the 1960s as a component of state-sanctioned, institutionalized racism. Reading test performance was used in some states as a registration requirement, thereby denying many African American citizens the right to vote (US Department of Justice, n.d.).

The measurement community began to turn concerted attention to defining, identifying, and removing unfairness in tests in the late 1960s and early 1970s as part of a larger societal movement to redress racial discrimination (Cole and Zieky, 2001). Similar concerns surfaced in the 1970s around accessibility and fairness for individuals with disabilities, most particularly with respect to postsecondary admissions tests (e.g., Sherman and Robinson, 1982; Willingham et al., 1988). Current concerns for the fairness and accessibility of tests for English language learners bring Brigham's (1930) statement full circle.

As noted, concerns for fairness are a social value, emerging first for fairness at the individual level and, later, for groups. Appreciation of the need for group fairness has been aided by the growing diversity of our society and the activism of those who were disenfranchised.

Concern for fairness will continue regardless of the form that future educational assessments take. Those tests will have to factor fairness into test design, delivery, scoring, analysis, and use. That concern will not be restricted to consequential tests but extend to formative assessment as well. Formative assessments entail a two-part validity argument: a) that the formative instrument or process produce meaningful inferences about what students know and can do, leading to sensible instructional adjustments and b) that these inferences and 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. Conceivably, a differentially valid formative assessment, used indiscriminately, could have the unwanted effect of *increasing* achievement gaps among population groups. Preventing such an occurrence might require the design and use of *demographically sensitive* formative assessments, in concept like pharmaceuticals created to

target particular population groups (Saul, 2005). In a free market system, however, development will be most concentrated on the needs of those most able to pay, leaving to government and advocacy organizations the task of ensuring that attempts are made to address instances of differential validity that disfavor underserved groups, when such instances do occur.

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 and 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. That is, test use was intended to focus educators in underachieving schools on the need to improve and, in particular, on improvement for underserved student groups.

Test design and use also can 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," in essence, an extreme curricular narrowing caused by the interaction of the Act's focus on reading and mathematics, a patchwork of mostly low-quality content standards among the states, the constrained methods used to measure achievement of those standards, and the sanctions placed on schools that fail to achieve required levels of proficiency.

The reasoning behind the *Race to the Top Assessment Program*, which the US Department of Education instituted to fund development of Common Core State Assessments, 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 a positive impact (US Department of Education, 2010). The implication of this claim is that impact must be explicitly taken into account at the assessment-design stage. By using principles and results from learning sciences research, summative assessments can be built to model good teaching and learning practice (Bennett, 2010). That modeling can occur via: a) giving students something substantive and reasonably realistic with which to reason, read, write, or do mathematics or science; b)

routinely including tools and representations similar to ones proficient performers employ in their domain practice; c) designing assessment tasks to help students (and teachers) connect qualitative understanding with formalism; d) structuring tests so that they demonstrate to teachers how complex performances might be scaffolded; and e) using learning progressions to denote and measure levels of qualitative change in student understanding.

Designing for positive impact might also mean preserving the idea of a consequential test — i.e., an event for which students must prepare. If the test is a faithful representation of the competencies and situations of use at which education is targeted, intensive preparation can have beneficial effects. Among other things, practice leads to automaticity, and to knowledge consolidation and organization. Testing can have positive effects by strengthening the representation of information retrieved during the test and also slowing the rate of forgetting (Rohrer and Pashler, 2010).

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. Assessment designers will also need to find new ways to enhance engagement. Designers might start by: a) posing problems that examinees are likely to care about; b) providing motivating feedback; c) using multimedia and other game elements; and d) employing delivery hardware preferred by the target population (e.g., smart phones, tablets), where that hardware is appropriate to the task demands of the domain.

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, either inside or outside of the game. 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 (relevant) competencies. That latter idea 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 and Mislevy, 1995).

In short, assessments of the future will need to be designed for engagement but not for the purpose of simply making assessments fun. Rather, they will need to be designed for engagement that facilitates, better than current assessments, measuring the competencies of interest for the assessment purposes at hand.

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 in generating information, certainly for the making of consequential decisions about individuals and institutions. Multiple sources are commonly are used for such consequential decisions as postsecondary admissions, where grade-point-average and tests scores are often are combined with one another through decision rules, and further clinically integrated with information from 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 for the class or an individual 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 other observations. Those other observations could be past classroom behavior, homework, quizzes, or the administration of additional tasks directly targeted at testing the hypothesis. As technology infuses learning and instruction, the amount and type of other information available only will increase.

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 all of the information needed for classroom, as well as for accountability purposes, will be gathered in the course of learning and instruction.

Whereas this idea may seem attractive on its surface, students (as well as teachers) have privacy rights that assessment designers will need to respect. For one, Individuals should know when they are being assessed and for what purposes. Their knowledgeable participation in assessment thereby becomes their informed consent. Second, having every learning (and teaching) action recorded and potentially used for consequential purposes is, arguably, an unnecessary invasion of the student's (and teacher's) right to engage freely in intellectual activity. That privacy invasion could potentially stifle experimentation in learning and teaching, including the productive making of mistakes (Kapur, 2010). Third, as a functionary of the state, the public school's right to ubiquitously monitor student and teacher behavior is debatable at best. In the US, at least, the state can monitor public behavior--as in the use of traffic and security cameras--particularly when that monitoring is in the interest of public safety. Except in very circumscribed instances, private behavior cannot be monitored without a court order. Whether the state can monitor learning behavior (as separate from testing behavior), and use that behavior to take actions that affect a student's life chances is an open question.

A compromise position that attempts to respect individual privacy and provide information for making consequential, as well as instructional, decisions might be a model similar to that used in many sports. In baseball, the consequential assessment of performance that counts toward player statistics and team standing occurs during the game, and only during the game. Spring training, before-game practice, in-between inning practice, and in between-game practice are primarily are reserved for learning. We might consider doing the same for assessment embedded in learning environments – use separately identified periods for consequential assessment versus learning (or practice).

Gather and Share Validity Evidence

However innovative, authentic, or engaging they may prove to 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 depends upon the assessment program providing honest evaluation, including independent analysis, 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 feeding validity results back into the improvement of the assessment program.

With respect to score generation, transparency must be apparent at least to members of the scientific community who are experts in the field, for it is these individuals who represent and advise the user community on technical matters. The need for transparency implies that score generation methods (e.g., automated scoring of constructed responses) cannot be so closely held by test vendors as to prevent independent review. In essence, "Trust us" approaches don't work when people's life chances are at stake.

One method for protecting intellectual property and permitting independent review is patent. A second, but less desirable approach from a transparency point of view, would be to grant access under a nondisclosure agreement to the user community's scientific advisors (e.g., members of a testing program's technical advisory committee). Those advisors could then report back to the user community in general terms that preserve the vendor's confidentiality but assure the technical quality of the scoring method.

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 (or at all) with traditional tests. Among those uses will be to measure existing competencies more effectively (and efficiently), for example, by scoring complex responses automatically or administering tests adaptively. A second use will be to measure new competencies. New competencies could include aspects of competencies we currently measure; for example, current tests measure the result of problem solving but technology also could be used to measure features of the examinee's problem-solving process (Bennett, Persky, Weiss, and Jenkins, 2010). Third, technology might be deployed to have positive impact on teaching and learning practice. Using technology without the promise of a clear substantive benefit ought to be avoided.

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.

Many of these claims are ones to which assessment programs have long aspired. However, meeting these claims in the face of an education system that will be digitized, personalized, and possibly gamified will require significantly adapting, and potentially reinventing, educational assessment. Our challenge as a field will be to retain and extend foundational principles, applying them in creative ways to meet the information and decision-making requirements of a dynamic world and the changing education systems that must prepare individuals to thrive in that world.

References

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

Bennett, R. E. (2005). *What does it mean to be a nonprofit educational measurement organization in the 21st century?* Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/Nonprofit.pdf

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., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*, 294-309.

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

Bennett, R. E., & Ward, W. C. (Eds). (1993). *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Erlbaum.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148. Retrieved from http://www.pdkintl.org/kappan/kbla9810.htm

Black, P., & Wiliam, D. (1998c). *Inside the black box: Raising standards through classroom assessment*. London, England: Kings College, London School of Education.

Brigham. C. C. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

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

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*, 1109–1136.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38,* 369-382.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York, NY: Teachers College-Columbia University.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. Philadelphia, PA: Center for Policy Research in Education.

Educational Testing Service (ETS). (2012). *The CBAL English Language Arts (ELA) Competency Model and Provisional Learning Progressions*. Princeton, NJ: Author. Retrieved from http://elalp.cbalwiki.ets.org/

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* (pp. 3-16). Lanham, MD: Rowan & Littlefield.

Kane, M. (2006). Validation. In R.Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport , CT: American Council on Education and Praeger.

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

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*,(4), 28–37.

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: American Council on Education and Praeger.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5–8.

Luecht, R.M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (Research Report 92-39). Princeton, NJ: Educational Testing Service.

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.

Miyazaki, I. (1976). China's examination hell: The civil service examinations of Imperial China. New York, NY: Weatherhill.

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.

Organisation for Economic Cooperation and Development. (n.d. a). Programme for International Student Assessment (PISA): PISA 2000 participants. Retrieved from http://www.oecd.org/pisa/participatingcountrieseconomies/pisa2000listofparticipatingcountrieseconomies.htm

Organisation for Economic Cooperation and Development (OECD). (n.d. b). Programme for International Student Assessment (PISA): PISA 2012 participants. Retrieved from http://www.oecd.org/pisa/participatingcountrieseconomies/pisa2012participants.htm

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 from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

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

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

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39,* 406-412.

Sherman, S. W., & Robinson, N. M. (Eds.). (1982). *Ability testing of handicapped people: Dilemma for government, science, and the public*. Washington, DC: National Academy Press.

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

Saul, S. (2005, June 24). F.D.A. Approves a Heart Drug for African-Americans. *New York Times*. Retrieved from http://www.nytimes.com/2005/06/24/health/24drugs.html

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

Shepard, L. A. (1991). Psychometricians beliefs about learning. *Educational Researcher, 20,* 2-16.

State of New Jersey, Department of Education. (n.d. b). *Guide to the New Jersey school report card 2011*. Trenton, NJ: Author. Retrieved from http://education.state.nj.us/rc/rc11/guide.htm

State of New Jersey, Department of Education. (n.d. a). *NCLB school, district, and state reports*. Trenton, NJ: Author. Retrieved from http://education.state.nj.us/rc/index.html

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

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H. I. Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Boston, MA: Allyn & Bacon.

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

US Department of Justice. (n.d.). *Introduction to federal voting rights laws*. Washington, DC: Author. Retrieved from http://epic.org/privacy/voting/register/intro_a.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/