Measuring What Matters: Part 1 – The Case for an Assessment Overhaul

by Jay McTighe

The emergence of the new Common Core Standards presents an opportunity to re-examine the current system of educational assessments and address their deficiencies. For the past ten years, The No Child Left Behind (NCLB) federal statute has required annual state testing as a means of gauging student achievement. Educational accountability under NCLB occurs as a result of publishing these test scores, comparing schools and districts, and enacting consequences for schools that fail to achieve "annual yearly progress" quotas. Responsible educators understand the need for accountability and the NCLB testing program has revealed achievement deficiencies that demand to be addressed. Nonetheless, the present assessment system is flawed, and ironically may impede the very efforts needed to attain important educational goals of the 21st century. In Part 1 of this article, I highlight several noteworthy deficiencies of current accountability assessments, and in Part 2, I will propose a more comprehensive assessment system that addresses these weaknesses and measures what matters most.

The adage, "what gets measured signals what is important," rings true in education. Students regularly ask their teachers, "will this be on the test?" If the answer is "no," they are less likely to pay attention to it. Large-scale assessments hold similar sway. Teachers and administrators pay close attention to what is tested on state and provincial assessments since their results can have high stakes consequences. If something is not assessed, it can quickly diminish in importance and receive less instructional emphasis. The adage applies to the current crop of assessments required by NCLB.

Currently, NCLB employs a "snapshot" approach to assessment through annual state testing in targeted subject areas. Given the large-scale nature of these tests, the majority of them understandably employ a selected-response format, allowing for inexpensive, machine scoring and relatively quick return of results. While multiple-choice tests provide broad, standardized measures yielding comparable results (at least within states), they are not well suited to assess certain key educational outcomes. For example, most state standards in English/Language Arts incorporate Listening and Speaking goals in addition to Reading and Writing. However, Listening and Speaking are not tested, and state assessments of writing vary in scope and quality.

Moreover, many subject areas for which standards exist are not tested at all in many states (e.g., history/social studies, science, visual and performing arts, technology).

To put it more starkly, important academic learning outcomes are falling through the cracks of the current large-scale assessment system. Selected-response assessments (or even brief-constructed responses) are simply incapable of measuring students' responses to open-ended problems and issues, discussion and debate, extended writing for real audiences and purposes, substantive research and experimental inquiry – yet these are surely vital outcomes. Furthermore, the so-called 21st Century Skills of creative thinking, collaborative teamwork, multi-media communication and use of information technologies are typically not tested on today's accountability measures. Accordingly, they are less likely to receive instructional emphasis. In sum, current standardized assessments capture what is easiest and inexpensive to test, but fail to assess many of the most valued goals of schooling.

High stakes assessments have consequential validity. In other words, their effects on curriculum, instruction, classroom assessments, and student motivation matter. For example, since repeated poor school performance on state measures can result in loss of accreditation, school reconstitution, and administrative transfers, educators (especially in low-achieving schools) are incentivized to focus on what is tested and disregard those standards (and even entire subjects) that are not. The result is often a *de facto* narrowing of the curriculum.

Furthermore, the pressure to improve performance on once-a year accountability assessments has prompted well-intentioned teachers and administrators to fixate on the format of the tests and institute a variety of misguided "test prep" interventions. While understandable, such actions reveal a misunderstanding -- the belief that the best way of improving accountability test scores is to practice the test format (multiple choice). Indeed, the emergence of "test prep" curricula and increased use of interim/benchmark assessments that mimic the state tests mistake the measures for the goals. Such practice is the educational equivalent of practicing for your physical exam in order to improve your health! Sadly, the use of classroom time in many schools (at least in the tested grades and subjects) would lead one to conclude that the Mission of schools is to improve test taking savvy and raise test scores rather than to strive for meaningful learning. Of course, it makes sense to familiarize students with test format, but excessive "multiple-choice" teaching and practice testing are *not* the best long-term strategies for developing a well-rounded, educated person *or* improving scores on yearly accountability tests.

Student motivation and engagement should not be overlooked when considering the impact of high stakes tests. Most learners are not stimulated by content "coverage," rote learning, skill drills, test prep exercises; and when students are bored by their schoolwork the consequences are well known – they exhibit a minimal-compliance attitude, they act up, or they drop out (figuratively and literally). A related casualty of the widespread use of multiple-choice practice tests and teacher-made assessments has to do with a worrisome lesson that this format suggests about learning; i.e., that the goal of school is to figure out the "correct" answer from a set of provided options.

How might a qualitative change to the assessment system address these shortcomings and negative effects of current high stakes measures? In the 2nd part of this article, I will propose a system that can minimize unhealthy curriculum narrowing, provide more robust evidence of academic knowledge as well as 21st century outcomes, and support meaningful learning through authentic and engaging instruction.

Measuring What Matters: Part 2 –
An Enhanced Assessment System Supporting Meaningful Learning

by Jay McTighe and Grant Wiggins

In "The Case for an Assessment Overhaul,[i]" Jay McTighe described deficiencies of the current assessment system. In Part 2 of "Measuring What Matters," we propose an assessment framework to address these deficiencies. The assessment framework we delineate offers an educationally viable approach for achieving three interrelated goals:

1) assessing the most important educational goals in appropriate ways;
2) providing the specific and timely feedback needed to improve learning; and
3) supporting curriculum planning, local assessment and instruction for meaningful learning.

In brief, we propose a "multiple measures" approach to educational accountability. Our framework consists of three inter-related components for assessing Core Standards and other important educational outcomes such as 21st Century Skills:

1) content-specific tests;
2) a series of content-specific and interdisciplinary performance tasks; and
3) a local assessment component.

This framework can be implemented nationally, through a consortium of states sharing the same items and tasks (i.e., components # 1 and 2), or on a state-by-state basis. In the event that states persist in using single, annual tests as currently specified by NCLB, our multi-measure assessment system can be modified for use at the district level. Each of the three assessment components is described below and Appendix A summarizes this proposed assessment system in chart form.

<u>Component #1</u> – Content-specific tests

The first component of is familiar to educators and the general public. It features content specific tests consisting of selected-response (SR) and brief constructed-response (BCR) items designed

to measure particular aspects of the Core Standards. Most current state tests and NAEP use SR and BCR items from which inferences about learning are drawn. These types of test have proven effective and efficient at sampling a broad array of basic knowledge and skills drawn from Standards. We recommend that these tests be computer-based in order to take advantage of enhanced item types made possible through technology-enabled assessments (for example, see Tucker, 2009[ii]), and to provide nearly immediate feedback in the form of detailed item analyses (not just scores). We further propose that a Matrix sampling approach be considered as a cost-saving means of obtaining accountability information at the school and district levels without subjecting every student to testing every year on every aspect of the Core Standards. However, states or school districts could opt for census testing if individual student scores are desired.

Component #2 – Content-specific and Interdisciplinary Performance Tasks

Selected-response and brief constructed-response item formats are limited in what they can appropriately assess. Performance tasks call for students to apply their learning to new situations in context. Accordingly, they are better suited to assess more complex aspects of Core Content Standards and Trans-disciplinary 21st Century Skills, such as mathematical reasoning, scientific investigation, issues analysis, creative problem solving, oral communications and technology applications.

The nation has a history of implementing performance assessments on a large scale. Current statewide writing assessments, The New Standards Project, and state assessments in Maryland, Connecticut, New York, California, Vermont and Kentucky conducted during the past two decades show the possibilities. Moreover, we have numerous district, state, and national models of judgment-based scoring of student work, including Advanced Placement, state and district-level writing assessments, music adjudications, and portfolio reviews in the visual arts. Other nations (e.g., Great Britain) include assessments scored by teachers as a major element of their national assessments.

The performance assessments would be set in real-world contexts and include both content-specific and interdisciplinary performances. We recommend that a national database of performance tasks and companion scoring rubrics be established from which national, regional or state assessments would be developed. In fact, many of these tasks and rubrics can be obtained from existing sets, such as those developed by The New Standards Project[iii] and as part of state assessments. Additional ones would be developed and certified by expert committees.

It is intended that the performance tasks be implemented by teachers *as part of the curriculum* at designated time periods during the school year. Scoring of the performance tasks will occur at regional scoring sites and be conducted by teams of teachers. State education departments and their regional services agencies would be responsible for the organization, training and monitoring of the scoring process to insure that consistent and reliable evaluation occurs. As a practical matter, schools and districts would be expected to align their academic calendars to the scoring schedule to ensure teacher participation during professional development days.

It is important to note that the scoring will *not* be contracted to commercial test companies, although companies may be enlisted to help with training, moderation and reporting. Indeed, a central feature of this proposal relates to the high-impact professional development that accrues when teachers work in teams to score student work. Accordingly, the costs of scoring the performance tasks need to be conceived and budgeted as a joint expenditure for assessment *and* professional development. An extension of the evaluation process occurs as teachers share ideas and resources for addressing the performance weaknesses observed during scoring. Emerging ideas for needed instructional interventions would be compiled in an Internet database, accessible to all teachers in the nation, region or state.

<u>Component #3</u> – Local Assessments

A standardized national or state assessment system is incapable of assessing each student on every important Standard and related educational goal (e.g., 21st Century Outcomes or the arts) for logistical and cost reasons. Even if it were feasible and affordable, it is unwise to limit accountability assessments to only those imposed from the outside. There is a need to include local assessments to allow appropriate measures of locally valued educational outcomes in *all* subject areas and to permit greater personalization than possible through external, standardized tests and tasks.

Standards are ultimately achieved at the local level. A comprehensive and effective national/state accountability system needs to include a district/school-level assessment component, and initiate policies and incentives to ensure that this local assessment becomes more credible, rigorous, and self-correcting. An analogy from athletics explains how this principle already works in the world of track and field. State officials do not have to officiate at every local track meet to be assured that the times and distances recorded by the local coaches are sufficiently accurate. There need only be local meets open to the public where the rules are followed and the scoring is transparent, backed

by a system of regional and state track meets – where local coaches need to worry about regional- and state-level performances, recorded by official scorers.

This third component of our assessment system is built upon the same logic; i.e., legitimize the role of local assessment by trusting educators with the responsibility of scoring student work in all subject areas. Make the results, framed in terms of Standards, public. Then, verify local scoring through a variety of regional and state audit systems.

The local component of the assessment system allows for a wide variety of possibilities, including common course exams, student projects and exhibitions, and interdisciplinary tasks involving collaboration and technology applications. More specifically, it:

• can appropriately assess important achievement targets (e.g., oral reading and speaking, applications of technology, collaborative teamwork) that may otherwise "fall through the cracks" of the first two components;

• is based on local curricula so that teachers, students and parents will be more likely to "own" the measures and the results;

• offers greater flexibility and potential for differentiation (e.g., giving students some choice of topics or products) than will the standardized assessments in the other components;

• honors the tradition of local control of education by allowing local decision making, rather than having all high-stakes assessments imposed from the outside;

• targets student accountability; i.e., the results become part of local grading and reporting (Thus, local report cards should have a section in which grades are provided on performance related to Core Standards and possibly 21st Century Skills.)

A cornerstone of this third component is a *Student Standards Folder* – a systematic collection of assessment evidence related to Core Standards and other important educational goals. The Standards Folder would:

- contain results from the **performance tasks** (described in Component #2);
- contain the results of **the content specific tests*** (described in Component #1);
- contain results from the **local assessments**;

- include **longitudinal (i.e., developmental) rubrics** in each subject area to guide judgments about student achievement *and* enable more systematic **tracking of growth** (i.e., progress toward meeting standards);*
- be **audited on an annual basis by regional-wide teams of educators** and citizen-experts, with two content areas sampled each year; and
- be **examined on a sampling basis by the state** in an audit of the quality of local and regional assessment.

*[Note: The test data would never be reported alone, but as a part of the overall Folder profile.]

Unlike a typical rubric used to evaluate student performance on a specific task or assignment, we recommend that the Standards Folders be judged against longitudinal rubrics based on developmental continua in various subject areas. For an example, see the American Council of the Teaching of Foreign Language Proficiency Guidelines.[iv] Such a system has been in place for over a decade in Great Britain for all subject areas[v]. Such rubrics enable educators, parents and students to track progress over time toward meeting exit standards.

The Standards Folder serves as a repository of a "body of evidence" of achievement *and* growth over time. Like a photo album, it provides a more complete and accurate portrayal of a learner than does any single test score ("snapshot"). It enables "triangulation" of data from multiple sources, ultimately yielding more credible (rich, varied, thorough) assessment evidence of Core Standards. Once in place, the Folder will enable students to graduate from high school with a *resume of accomplishment* compiled over their school career, rather than simply a transcript of courses taken, "seat time" logged, and a cumulative GPA.

In summary, we contend that the proposed 3-part system provides a more comprehensive assessment of Core Standards, while avoiding many of the problems of current NCLB accountability testing.

[i] McTighe, J. "Measuring What Matters: Part 1 – The Case for an Assessment Overhaul" in *What's Working in Schools Newsletter*, December 2010. Bloomington, IN: The Hope Foundation.

[ii] Tucker, B. "Beyond the Bubble: Technology and the Future of Student Assessment." *Education Sector Reports.* February, 2009.

URL < http://www.educationsector.org/publications/beyond-bubble-technology-and-future-student-assessment>

[iii] New Standards Project. Washington, DC: National Center on Education and the Economy.

URL <http://www.ncee.org/about-ncee/history/>

[iv] American Council for the Teaching of Foreign Languages. 1983. *ACTFL Proficiency Guidelines.* Revised 1985. Hastings-on-Hudson, NY: ACTFL Materials Center.

URL <http://www.sil.org/lingualinks/languagelearning/otherresources/actflproficiencyguidelines/contents.htm>

[v] Click on "Assessment of Subjects", then "Progressions" to view the developmental rubrics

URL <http://curriculum.qcda.gov.uk/key-stages-1-and-2/assessment/assessmentofsubjects/assessmentinartanddesign/index.aspx>

Assessment Component	Potential Benefits	Potential Drawbacks	Costs
1. Content-specific Standardized Tests • selected-response and brief constructed response formats • generally de-contextualized	• able to sample a broad array of knowledge and skills within Core Standard areas • quick and inexpensive scoring and reporting • familiar test format • items can be drawn from existing banks (e.g., state tests, NAEP, NWEA) • allows for computerized testing • standardization allows for comparable results • can be used for school/ dirict accountability	• can encourage de-contextualized "test prep" at the expense of meaningful learning • may lead to a narrowing of the curriculum (i.e., focus only on the tested content) • cannot fully measure important learning areas (e.g., mathematical reasoning, critical thinking, extended writing, research) • tests are generally not known in advance	• comparable to current standardized testing programs* *A national testing program ('ala NAEP) would be more cost-effective than mounting 50 different state programs. * A matrix-sampling model could be used to reduce costs (but at the expense of providing individual student scores on every test).

2. Content-specific and Interdisciplinary Performance Tasks • open-ended • require extended constructed responses • allow for contextualized and authentic application • tasks are scored at regional scoring sites by practicing teachers • require rubrics, anchors and inter-rater protocols for reliable scoring	• able to provide more valid measures of important learning (e.g., mathematical reasoning, critical thinking, extended writing) in greater depth • able to assess learners' understanding through contextualized (i.e., more genuine) application, including interdisciplinary contexts • 21st Century Outcomes (e.g., technology use, collaborative skills) can be integrated with academic knowledge • tasks can be drawn from existing banks (e.g., states [MD, KT, CT] and The New Standards Project) • "practicing" for the tasks can support meaningful learning • more transparent (i.e., basic tasks and scoring rubrics are known) • standardized rubrics and scoring procedures allow for comparable results • significant professional learning can result for teachers involved in the scoring • can be used for school/district accountability	• less able to measure a breadth of knowledge and skills • time-consuming to give and score • expensive to score • judgment-based scoring may compromise reliability • delayed results due to time required for scoring	• Cost estimates can be obtained from several states (MD, CT, KY) that have implemented large-scale performance assessment programs, as well as from many more that conduct statewide writing assessments. * The costs of scoring the performance tasks should be viewed as expenditures for both measurement *and* professional development of teachers.

3. Local Assessments • allow for a variety of assessment types (e.g., course exams, Senior projects, portfolio collections) • based on local curricula • can be used for student accountability and local grading • features a Student Standards Folder to serve as a repository of achievement evidence • scored against developmental continua (longitudinal rubrics) • not standardized outside of a school or district, so cannot be used for state, district or national comparisons	• allow for a variety of assessment types (e.g., course exams, Senior projects, portfolio collections) aligned to local curricula • promote local options and greater "ownership" of measures and results • allow for assessing important learning goals that otherwise "fall through the cracks" of the standardized assessments (# 1 and 2) • provide more immediate and credible feedback • encourage curriculum fidelity and focused instruction • can allow for differentiation and student choice (e.g., on products) • yield individual student scores; can be used for student accountability (e.g., grading) • track progress along developmental continua toward meeting standards	• results are not comparable beyond the school or district • not suitable for use in school/district accountability	• Costs would be dependent on the nature of the curriculum and the chosen assessment options. In general, these costs would be assumed by the local school/district budget.