FairTest
National Center for Fair & Open Testing

Multiple Measures:
A Definition and Examples from the U.S. and Other Nations

Summary

Definition. Multiple measures: the use of multiple indicators and sources of evidence of student learning, of varying kinds, gathered at multiple points in time, within and across subject areas.

Examples of multiple measures systems used successfully in the U.S.

- *Learning Record*: Developed for use with multi-lingual, multi-cultural populations, to assess progress in reading, writing, speaking and listening. Using a structured format, the teacher regularly observes and describes the student and her work, and attaches samples, to provide multiples sources of evidence. Student progress is summarized in writing and placed numerically on a developmental scale. LRs have been re-scored with high inter-rater agreement, and studies have supported its validity.

Work Sampling System (WSS): The WSS, designed for students aged 3-8, facilitates the collection and evaluation of observations and examples of student work. Learning is summarized in writing and numerically. It was demonstrated to have strong validity and good reliability.

- *New York Performance Standards Consortium*: 26 NY high schools have a variance to use one state exam (ELA) out of a mandated five. The Consortium uses the ELA test and four consortium-wide performance tasks for graduation. The test and the math task are used to determine adequate yearly progress under NCLB. Student work is evaluated by teachers and independent reviewers. The system has been reviewed and approved by independent experts.

- *Nebraska STARS*: A statewide system of local assessments, independently reviewed and approved as meeting state standards for assessment, then periodically audited for quality. Types of assessments varied across districts, but most developed multiple measures and performance tasks. Independent reviews affirmed assessment quality. Districts administered state writing exams and norm-referenced tests in three grades. The system met federal resistance after NCLB and was replaced by a statewide test.

- *Wyoming's* "Body of Evidence" approach uses locally developed assessments, incorporating multiple measures, designed to indicate students have met state graduation standards. The local assessment systems are evaluated through a peer-review process.

Examples from other nations. Most other nations, including many with better outcomes on various indicators, test less than the U.S. They use a mix of state/national and local assessments, including performance tasks, primarily for public information and improvement efforts, not accountability.

- *Queensland, Australia, "Rich Tasks"*: In a pilot program, extended multi-disciplinary performance tasks of varying types, for use in three grades, were developed centrally, integrated with the local curriculum, and used when teachers decided. Teachers judged student performance against pre-set standards. Queensland used a "moderation" process in which teams of teachers re-scored

samples of student work. Rich Task scores were included in student grades. The formal pilot has ended, but the tasks are still used in many schools, with some state support.

Queensland: Aside from the "New Basics/Rich Tasks" project, Queensland uses multiple forms of assessment and relies on local assessments. Linda Darling-Hammond explains:

"Until the early 1970s, a traditional "post-colonial" examination system controlled the curriculum. When it was eliminated, all assessments became school-based. Teachers develop, administer, and score the assessments in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and panels that include teachers from other schools as well as at least one professor from the tertiary education system moderate the assessments." [Darling-Hammond is the source of the other quotations in this section.]

Finland: "Finland has no external standardized tests used to rank students or schools… Finland's leaders point to its use of school-based, student-centered, open-ended tasks embedded in the curriculum as an important reason for the nation's extraordinary success on international exams… School-level samples of student performance are evaluated periodically by the Finnish education authorities, generally at the end of the 2nd and 9th grades, to inform curriculum and school investments. All other assessments are designed and managed locally."

Sweden "pairs its nationally outlined and locally implemented curriculum with multiple layers of assessment controlled by schools and teachers. Assessments in compulsory school consist of several components… Teachers keep extensive records of student progress, using three assessments to aid in their grading at the Upper Secondary school level: 1) coursework, 2) assessments designed by teachers based on the course syllabi, and 3) nationally approved examinations when grading the core subjects… Regional education officials and schools provide time for teachers to calibrate their grading practices to minimize variation across the schools and across the region."

Hong Kong's "assessment system is evolving from a highly centralized examination system to one that increasingly emphasizes school-based, formative assessments that expect students to analyze issues and solve problems." In some high school examinations, 20-30% of the grade is derived from classroom-based performance tasks.

Singapore's system is evolving toward greater use not only of performance tasks, but also school-based evidence. Exams count in college entry decisions, but not for graduation. Some high school tests include school-based components. The education system encourages multiple forms of assessment in earlier grades as well. However, this information is not part of a larger assessment system since such a system does not exist prior to exams at the end of primary school (year 6, age 12). These national exams "are administered and scored by teachers in moderated scoring sessions."

United Kingdom. England uses multiple measures, both in-school and in the combination of school-based and external assessments used for accountability. Teacher judgments are moderated at the school or national level, depending on which grade ("key level"). Wales has eliminated national exams for children through age 14. Teachers create and score assessments prior to the college entry exams. Northern Ireland "is in the process of implementing an approach at all levels called 'Assessment for Learning.' This approach emphasizes locally developed, administered and scored assessments." There are no mandated government tests through age 14.

International Baccalaureate. "[T]eachers conduct school-based assessments by grading individual pieces of coursework based on the objective set out by the IB subject outlines. School-based assessments contribute between 20 and 30% of the total grade in most subjects," and more in others.

Conclusion: Multiple measures, extensive use of performance assessments and the inclusion of local evidence are feasible in large-scale assessment systems. Through reviews of such systems, using auditing and moderation, both reliability and comparability can be established.

Multiple Measures:
A Definition and Examples from the U.S. and Other Nations

Monty Neill, June 2, 2010

I. Definition.

Multiple measures: the use of multiple indicators and sources of evidence of student learning, of varying kinds, gathered at multiple points in time, within and across subject areas. These include but are not limited to: teacher observations; tests that include multiple-choice, short and longer constructed response items; essays; tasks and projects of various sorts done in various modes including electronic; laboratory work; presentations; and portfolios. They are used to assess higher-order thinking skills and understanding, including analysis, synthesis, evaluation, application, problem-solving and creativity. They are used for both formative and summative purposes, and many become part of the learning process itself; we can thus speak of assessment *for*, *as* and *of* learning.

- See Appendix A for further elaboration on this definition.

II. Examples of multiple measures from the United States.

The demands of No Child Left Behind (NCLB) have stymied the use of multiple measures in the U.S. However, important work has been done in the U.S., and some continues and even grows.

A. The Learning Record

The *Learning Record* (LR) was developed in England as the *Primary Language Record* for use with multi-lingual, multi-cultural populations in reading, writing, speaking and listening. Its structure provides a consistent framework for gathering and evaluating information. It was adapted and expanded in the U.S. and was beginning to grow, particularly in many Bureau of Indian Affairs schools, before being largely swept aside by NCLB requirements.

In the LR, each teacher documents and evaluates student work and progress, focusing on reading and writing. Thus, a student's LR would include documentation on books the student can read and understand, including evidence showing her understanding of them, as well as samples of writing and teacher observations of the student as a learner. Since the specific books each child reads vary, the specific evidence varies by student. The LR includes a variety of types of evidence as well. Each student's progress is documented, summarized in writing, and placed numerically on a developmental scale.

The LR scales have been validated. Moderation processes (independent review of Records) have established adequate to superior inter-rater agreement between moderators and the teachers providing the initial scores. This shows that with a good structure, diverse sources of information can be brought to bear on common topics (e.g., reading development). It also supports the accuracy of teachers' judgments in placing their students on the developmental scales. Such judgments (numbers) can be aggregated and used to describe overall student attainment and progress. That is, if each originating teacher's judgment is sound (supported by a review of 3-5 randomly sampled Records), then the aggregate information about classrooms and schools can be considered sound.

LR practice and research demonstrate it is a reliable, valid, comparable and educationally sound method of evaluating individual progress and status using multiple sources of evidence, and of aggregating that information to provide public information about schools.

- For more information, see http://www.fairtest.org/learning-record.

B. Work Sampling System.

The *Work Sampling System* (WSS) for children aged 3-8 is similar in some ways to the LR. It was developed by Samuel Meisels, one of the nation's foremost authorities on the assessment of young children. It includes collecting, evaluating and summarizing observations and examples of student work. It too was demonstrated to have strong validity and reliability.

The WSS uses three complementary elements to assess student knowledge and development: 1) observations by teachers using Developmental Guidelines and Checklists, 2) collection of children's work in Portfolios, and 3) Summary Reports. The Developmental Guidelines used in teacher observations are based on national content standards and current knowledge of child development. This gives all observations the same basis of description and evaluation. Observations and the collection of materials for portfolios continue throughout the school year. Summary Reports are produced and distributed to parents and students three times a year fall, winter and spring.

The WSS is now owned by Pearson; information on it is available at http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=PAworksampl&Mode=summary. See also, http://www.fairtest.org/work-sampling-system; and http://www.fairtest.org/trusting-teacher-judgment.

C. New York Performance Standards Consortium

The Consortium includes 28 public high schools, most in New York City. They have received a variance from the state that allows their schools to use only one of five mandated Regents exams (Language Arts) and instead use their combination of consortium- and school-based performance assessments for both ESEA's AYP and for graduation requirements. The Consortium's website reports:

"Consortium schools have devised a system of assessment which consists of eight components including alignment with state standards, professional development, external review, and formative and summative data. Consortium schools have documented how their work meets and exceeds New York State Regents standards through a system of rigorous commencement-level performance-based assessment tasks. Performance on these tasks is reflected on student transcripts and results are used for college admission.

"The tasks require students to demonstrate accomplishment in analytic thinking, reading comprehension, research writing skills, the application of mathematical computation and problem-solving skills, computer technology, the utilization of the scientific method in undertaking science research, appreciation of and performance skills in the arts, service learning and school to career skills. Experts external to the schools, from universities and the business world, participate in reviews of student work.

"The Performance Assessment Review Board, Inc., an external body of educators, test experts, researchers and members of the legal and business world, monitors the performance-based assessment system and systematically samples student work."

Consortium members use performance assessments instead of Regents tests in a series of graduation-level tasks: analytical comparative essay in literature, social studies research paper, original science experiment, and application of higher level mathematics, as well as proficiencies in oral defenses and exhibitions of their work.

While the Consortium relies on performance assessments, these include a variety of kinds of tasks and projects over the various subjects, providing multiple opportunities for students to demonstrate their learning in different ways over time. It is therefore a multiple measures system.

Common rubrics are used to help ensure consistency in scoring. The Consortium validated the use of the rubrics in four subjects through shared re-scoring of sample work.

- See Appendix B for a map of the structure of the Consortium.
- For more information, see http://performanceassessment.org/consortium/index.html.

D. Nebraska STARS

Prior to NCLB, Nebraska developed its Statewide Teacher-led Assessment and Reporting System (STARS), composed of local assessments that met statewide standards, including that they be based on the state's academic content standards (or state-approved equivalent local standards), ensure consistent scoring, be unbiased and developmentally appropriate with mastery levels set appropriately, and that students must have an opportunity to learn the content. In most districts, local educators helped develop the assessments. Each local system was reviewed by independent experts. If it was not approved, it was revised until it met the criteria. The state then audited the district assessment system periodically and if a district proposed major changes.

The nature of the assessments varied. Districts often used criterion-referenced tests standardized within the district. Most incorporated more extended, classroom-based work, including tasks and projects. These had a positive impact on teaching and learning. The state system included a state-wide writing examination and the administration of a norm-referenced test at three grades. While these were not used for accountability, they served as a check on the validity of the local assessments. Independent reviews, such as by the Buros Institute for Mental Measurements, found the districts generally produced strong assessments and were willing to improve.

Though U.S. Education Secretary Rod Paige had expressed support for Nebraska's system, his successor, Margaret Spellings, opposed the state's efforts, blocked approval, and worked with legislators in the state to switch to a single-test model.

For more on Nebraska see:

- *Reclaiming Assessment: A Better Alternative to the Accountability Agenda*, Chris Gallagher, 2007, Heinemann;

- articles in *Phi Delta Kappan*: "Turning the Accountability Tables: Ten Progressive Lessons from One 'Backward' State," by Chris Gallagher, January 2004; "Nebraska STARS: Achieving Results," by Pat Roschewski, Jody Isernhagen, and Leon Dappen, February 2006; and

- a search for Nebraska at www.fairtest.org will turn up articles about the system and the political battle to save it.

E. Wyoming's "Body of Evidence"

Wyoming has implemented a "Body of Evidence" approach as part of its high school graduation requirements. These are locally developed assessments designed to indicate students have met state graduation standards. The state's website says, "The philosophy at the heart of the Wyoming Body of Evidence system is to provide multiple measures to assess student mastery of the content standards; in this way, no single assessment can disqualify a student from graduation."

The Body of Evidence is a collection of a student's work proving understanding of concepts and the ability to perform certain required skills. In keeping with the emphasis on locally designed approaches, Wyoming allows four different ways for a district to design a Body of Evidence, and districts can choose the way or combination of ways that best suits their needs. As in Nebraska, a district's assessment must meet specified criteria:

- Provide evidence of student achievement directly related to the Wyoming state standards.
- Give students multiple opportunities and multiple ways (i.e., not just more chances to take the same tests) to demonstrate their knowledge and skills relating to the standards.
- Be fair to all students, including those with disabilities or who are learning English, and provide accommodations.
- Allow education professionals to decide what's "good enough" in a fair and reasonable way.
- Create assessments that are similar across schools and classrooms within the same school district both within a given year and across years.
- Answer these two questions: Does the student know enough to graduate? And does the evidence support the answer?

The state website says, "Each district's Body of Evidence system is reviewed through a peer review process facilitated by the Wyoming Department of Education."

- For more information see http://www.k12.wy.us/SA/BOE.asp and http://www.fairtest.org/wyoming-steers-clear-exit-exams.

F. Other states with potentially useful components

Multiple Measures Approaches to High School Graduation, by Linda Darling-Hammond, et al., describes various approaches to graduation assessments (both mandated and voluntary) in place in various states. (The School Redesign Network at Stanford University, 2005; available at http://www.srnleads.org/data/pdfs/multiple_measures.pdf .) Other, older information can be found at FairTest: *Annotated Bibliography: Performance Assessment* (http://www.fairtest.org/annotated-bibliography-performance-assessment) and *Implementing Performance Assessments*, by Monty Neill, et al.

III. Examples from Other Nations

Linda Darling-Hammond (2010) summarizes how other nations are using multiple measures with a focus on performance tasks:

> "Whereas U.S. tests rely primarily on multiple-choice items that evaluate recall and recognition of discrete facts, most high-achieving countries primarily rely on open-ended items that require students to analyze, apply knowledge, and write extensively. Furthermore, these nations' growing emphasis on project-based, inquiry-oriented learning has prompted increased use of school-based tasks, which include research projects, science investigations, development of products, and related reports or presentations. These assessments, which are incorporated into the overall examination scoring system, help focus the day-to-day work of teaching and learning on the development of higher-order skills and use of knowledge to solve problems… In many cases, school-based assessments complement centralized 'on-demand' tests and may constitute up to 60% of the final examination score… [Tasks] are generally designed, administered, and scored locally, based on common specifications and evaluation criteria… decisions about when to undertake these tasks are made at the classroom level, so they are used when appropriate for students' learning process." [Linda Darling Hammond (2010), *Benchmarking Learning Systems: Student Performance Assessment in International Context*; available at http://edpolicy.stanford.edu/pages/pubs/pub_docs/assessment/scope_pa_ldh.pdf; unless noted otherwise, quotations in this section are taken from this paper, with the author's permission.]

Across the nations, prior to college entry testing there are varying amounts of national/state exams, ranging from none to three. These typically utilize multiple formats. The stakes in most of these nations are low or none for students prior to college entrance exams, which are commonly a part of making admissions decisions. Stakes also are low or none for schools. Most assessing is classroom- or school-based, though often based on national or state curriculum frameworks. It is normal practice to employ a variety of kinds of measures. Where performance tasks and projects predominate, they incorporate a variety of such assessments.

Though these assessments are mostly not high-stakes, it is instructive to see how other nations include them in their educational systems, particularly the employment of auditing and moderation as means to ensure comparability across schools. The use of multiple measures for college admissions further demonstrates that other nations are successfully incorporating multiple sources of evidence to make decisions about students. Darling-Hammond also contextualizes these assessment systems with brief descriptions of history, the role and nature of standards, etc.

- Darling-Hammond's paper contains a useful summary table that includes, for each country, a description of the assessment system, "What kinds of assessments are used?" and "Who designs and grades assessments?" It is on pp. 39-44.

A. Queensland, Australia, "New Basics" and "Rich Tasks" Project

The state's "New Basics" and "Rich Tasks" approach to standards and assessment, which began as a pilot in 2003, offers extended multi-disciplinary tasks that are developed centrally and used locally when teachers decide the class is ready. They can be integrated with locally-oriented curriculum. They are, says Queensland's reports, "specific activities that students undertake that have real-world value

and use, and through which students are able to display their grasp and use of important ideas and skills." Rich tasks enable identification of "mandated student knowledge, skills and practice outcomes at critical junctures of schooling." Their use provides "conditions for local school-specific curriculum development in response to community needs." Students do a range of diverse tasks, thereby providing multiple measures.

Extensively researched, this system has had success as a tool for school improvement. Studies found stronger student engagement in learning in schools using the Rich Tasks. On traditional tests, New Basics students scored about the same as students in the traditional program, but they performed notably better on assessments of higher order thinking. Scores for lower achieving students in New Basics schools rose, and gaps between White and Aboriginal students narrowed.

In scoring the Rich Tasks, teachers judge student performance against pre-set standards. The expectation is that schools will learn to grade samples of Rich Tasks consistent with statewide standards. All tasks are scored by teachers within a school (internal moderation). A sample is re-scored at the district level (external moderation), with feedback to the school leading to possible changes to student scores. Separate samples are centrally reviewed to compare schools and scoring across the state.

Students have a stake in the outcome since their grades include scores from Rich Tasks. There are no mandated high-stakes uses of the results for schools.

Independent reviewers determined that the tasks and the scoring processes were valid and reliable. The reviewers concluded the processes "provide the necessary assurance and information to the school to inform parents that the grades, achieved by their child, are consistent and comparable with state-wide standards."

As in *every* situation in which different tests taken in different years must be equated, or when tests at different grades are put on a single scale, there are technical complications: "There is a process of statistical adjustment to place the Year 5 and 7 scores on the same scale as the Year 3 scores within a calendar year, and to place the current year's scores on the same scale as the previous year's. However, this is a process of estimation that introduces cumulative errors and a loss of precision." This issue also affects traditional U.S. standardized tests.

- For more information, see http://education.qld.gov.au/corporate/newbasics/, which has sample tasks and many research papers used in writing this summary.

B. Queensland, Australia

Aside from the New Basics/Rich Tasks project, Queensland uses multiple forms of assessment and relies on local assessments:

"Until the early 1970s, a traditional "post-colonial" examination system controlled the curriculum. When it was eliminated, all assessments became school-based. Teachers develop, administer, and score the assessments in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and panels that include teachers from other schools as well as at least one professor from the tertiary education system moderate the assessments."

Darling-Hammond describes the state's assessment process for physics, which includes extended tasks and a variety of other work:

"At the end of the year, teachers collect a portfolio of each student's work, which includes the specific assessment tasks, and grade it on a 5-point grading scale. To calibrate these grades, teachers put together a selection of portfolios from each grade level—one from each of the 5 score levels plus borderline cases—and send these to a regional panel for moderation. The panel of five teachers re-scores the portfolios and confers about whether the grade is warranted, making a judgment on the spread. State review panels also look at a sample of student work from each district to insure that schools implement the standards across all districts. Based on this analysis and a 12th grade standardized state-wide test called the Queensland Core Skills (QCS) Test, the Queensland authority confirms the levels of achievement proposed by school programs and may adjust it if it does not calibrate to the standards."

Note: Darling-Hammond also describes the state of Victoria, which also uses multiple forms of evidence, including classroom-based assessments.

C. Finland

"Finland has no external standardized tests used to rank students or schools… Finland's leaders point to its use of school-based, student-centered, open-ended tasks embedded in the curriculum as an important reason for the nation's extraordinary success on international exams… School-level samples of student performance are evaluated periodically by the Finnish education authorities, generally at the end of the 2nd and 9th grades, to inform curriculum and school investments. All other assessments are designed and managed locally. The national core curriculum provides teachers with recommended assessment criteria for specific grades in each subject and in the overall final assessment of student progress each year… Local schools and teachers then use those guidelines to craft a more detailed curriculum and set of learning outcomes at each school as well as approaches to assessing benchmarks in the curriculum… Teachers are treated as "pedagogical experts" who have extensive decision-making authority in the areas of curriculum and assessment in addition to other areas of school policy and management… Teachers' reports must be based on multiple forms of assessment, not only exams." The core curriculum is very brief in each subject; the one for math is only 10 pages for all the grades together.

College entry exams are constructed by college professors and high school teachers, and are scored by local teachers. "[S]amples of the grades are re-examined by professional raters."

D. Sweden

"Over the past 40 years, Sweden's national assessment system has, like Finland's, shifted from a centralized system based on one test to a more localized system based on multiple forms of assessments… Sweden pairs its nationally outlined and locally implemented curriculum with multiple layers of assessment controlled by schools and teachers. Assessments in compulsory school consist of several components…

"[S]tudents take nationally approved examinations in year 9. The exams assess the subjects of Swedish, Swedish as a second language, English, and mathematics. Teachers use these assessments as one factor in determining students' grades. The exam at year 9 is compulsory for schools, but not for students. Sweden uses the scores from the test to ensure the grades given by teachers compare to the national standards…

"Teachers keep extensive records of student progress, using three assessments to aid in their grading at the Upper Secondary school level: 1) coursework, 2) assessments designed by teachers based on the course syllabi, and 3) nationally approved examinations when grading the core subjects of Swedish, English and mathematics, and selected other areas… Regional education officials and schools provide time for teachers to calibrate their grading practices to minimize variation across the schools and across the region…

"The National School Board examinations administered during Compulsory and Upper Secondary schooling use an open-ended, authentic approach to assessing students…"

E. Hong Kong

"Hong Kong's assessment system is evolving from a highly centralized examination system to one that increasingly emphasizes school-based, formative assessments that expect students to analyze issues and solve problems… [S]chool-based assessments….are assuming greater prominence in the government's plan to… combine on-demand tests with curriculum-embedded tasks." Some now in existence have 20-30% of the grade derived from school-based assessments.

"The Hong Kong Education Examinations Authority explains the rationale for growing use of school-based assessments (SBA) [in its high school exit requirements]:

> 'The primary rationale for SBA is to enhance the validity of the assessment, by including the assessment of outcomes that cannot be readily assessed within the context of a one-off public examination. SBA can also reduce dependence on the result of public examinations, which may not always provide the most reliable indication of the actual abilities of candidates. Obtaining assessments based on student performance over an extended period of time and developed by those who know the students best - their subject teachers - provides a more reliable assessment of each student…
>
> 'Teachers know that SBA, which typically involves students in activities such as making oral presentations, developing a portfolio of work, undertaking fieldwork, carrying out an investigation, doing practical laboratory work or completing a design project, help students to acquire important skills, knowledge and work habits that cannot readily be assessed or promoted through paper-and-pencil testing.'

"The Education Bureau… promotes the use of multiple forms of assessment in schools including projects, portfolios, observations, and examinations, and looks for the variety of assessments in the performance indicators used for school evaluation."

F. Singapore

The Singapore system is evolving toward greater use not only of performance tasks, but also school-based evidence. Exams count in college entry decisions, but not for graduation.

The Education system encourages multiple forms of assessment in earlier grades as well. However, this information is not part of a larger assessment system since such a system does not exist prior to exams at the end of primary school (year 6, age 12). "At the end of Year 6 (age 12), students take the Primary School Leaving Examinations (PSLE). These are open-ended written and oral examinations… that are administered and scored by teachers in moderated scoring sessions." There also are national exams at the end of year 10.

"Students attending Junior College (grades 11 and 12) en route to university take the GCA Advanced Level (A-Level) exams at the end of year 12 (age 18)… . A number of the high school content tests are accompanied by school-based tasks, such as research projects and experiments designed and conducted by students… These school-based components, which teachers manage and score according to specifications provided by the Examinations Board, count for up to 20% of the examination grade. Scoring is both internally and externally moderated" (that is, moderation occurs at two stages, within a school and across schools; the Learning Record employed this process in many schools.) Pre-university students must also do an extended, multi-component project (e.g., with work products and oral presentation; individual and group work); this work also is both internally and externally moderated.

G. United Kingdom

England

"Teachers assess pupils' progress continuously and assemble evidence for external reporting in the national data system at ages 7, 11, and 14 (Key Stages 1, 2, and 3). This evidence is based on classroom-based assignments, observations, and tasks, the results of which are evaluated in terms of indicators of performance outlined in learning progressions for each of several dimensions of learning within each subject area."

England uses multiple measures, both in-school and in the combination of school-based and external assessments:

"At Key Stage 1, student progress is evaluated based on classroom evidence and results from centrally-developed, open-ended tests and tasks in English and mathematics. The tests and tasks are marked by teachers and moderated within the school and by external moderators. At Key Stage 2, student progress is evaluated based on teachers' summary judgments and results from open-ended tests in English, mathematics, and science. These tests are externally marked and the results reported on a national level. For Key Stage 3, England recently abolished external tests and now relies on teacher assessments to report achievement levels in all subjects. Teacher judgments are moderated and results are reported on a national level." Teacher and parent opposition to these tests fueled their elimination. Teachers this year have largely boycotted the administration of Stage 2 tests.

"At Key Stage 4, ages 15 to 16, the national qualification framework includes multiple pathways for students and consequently multiple measures of student achievement… Most students take the GCSE, a two-year course of study evaluated by assessments both within and at the end of courses or unit… [Students can select how many and which exams to take.] The exams involve constructed response items and structured, extended classroom-based tasks which comprise from 25 to 60% of the final examination score."

Wales

"Wales broke from the British system and opted to abolish national exams for children through age 14… Much like Finland, during the primary years Welsh schools have a national school curriculum supported by teacher-created, administered, and scored assessments. During the secondary years, teachers create and manage all assessment of 14-year-old students, while students 16 years and older are encouraged to participate in the relevant [national] GCSE exams and A-level courses and exams."

FairTest's investigation into what happened when Wales dropped the tests reported: "What do Welsh teachers use instead of the tests? With government guidance, teachers come up with their own

assessments and report the results to parents, local education authorities, and the Welsh government each year. Freed from the need to prepare students for narrow tests, secondary school teachers employ out-of-school experiences, in-depth research, and presentations, emphasizing applied learning in secondary school and underscoring the importance of play in early childhood education." (http://www.fairtest.org/wales-drops-most-standardized-testing.)

Northern Ireland

"Northern Ireland is in the process of implementing an approach at all levels called "Assessment for Learning." This approach emphasizes locally developed, administered and scored assessments.

"Northern Ireland does not require schools to externally assess students up through age 14, but it provides teachers with the option to give students assessments at the end of Stage 3, which are externally graded."

H. International Baccalaureate (IB)

"In almost all of the subjects, teachers conduct school-based assessments by grading individual pieces of coursework based on the objective set out by the IB subject outlines. School-based assessments contribute between 20 and 30% of the total grade in most subjects and as much as 50% in arts courses like music, theater arts, and visual arts. Coursework graded by teachers includes such assessments as oral exercises in language subjects, projects, student portfolios, class presentations, practical laboratory work, mathematical investigations, and artistic performances…

"There are a limited number of externally assessed pieces of work (i.e., a theory of knowledge essay, extended essay, and world literature assignment) that students complete over an extended period of time under teacher supervision, but which are marked by external evaluators, or 'IB Examiners,' personnel trained and organized by the IBO."

IV. Bibliographic note

For another discussion on using multiple forms of assessment, including local and performance, see Wood, George H., Linda Darling-Hammond, Monty Neill, and Pat Roschewski. 2008. "Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills." Available at http://www.fairtest.org/refocusing-accountability.

V. Appendices

A. Elaboration of the definition of "multiple measures"

The Forum on Educational Accountability (FEA) calls for the use of multiple measures in evaluating students, educators, schools and districts. FEA's 2007 Assessment report included the following:

"Coherent and comprehensive assessment systems provide evidence of student and school performance in relation to rich and challenging educational goals, using multiple indicators of student learning from a variety of sources at multiple points in time… Comprehensive assessment systems would address these areas through employing multiple appropriate assessment practices and tools, including: teacher observations; tests that include multiple-choice, short and longer constructed response items; essays; tasks and projects; laboratory work; presentations; and portfolios. It would also include development of assessments for specific subgroups, including English language learners (ELLs) and students with disabilities (SWDs)." [Principle II; report available at http://www.edaccountability.org/AssessmentFullReportJUNE07.pdf]

FEA's 2007 recommendations for the reauthorization of the Elementary and Secondary Education Act, FEA drafted legislative language that elaborated the definition as amendments to Sec. 1111(b)(3) Academic Assessments:

(vi) {Replace current language with:} involve multiple up-to-date assessments of student academic achievement, including assessments that assess higher-order thinking skills and understanding, including analysis, synthesis, evaluation, application, problem-solving and creativity, in and across subject areas.

(I) Multiple assessments involve different sources and kinds of evidence of student learning in a subject or across subject areas.

(II) They may include state-level assessments; classroom, school and district tests; extended writing samples administered on demand or as part of classroom work; tasks, projects, performances, and exhibitions; and collected samples of student classroom work, portfolios or learning records.

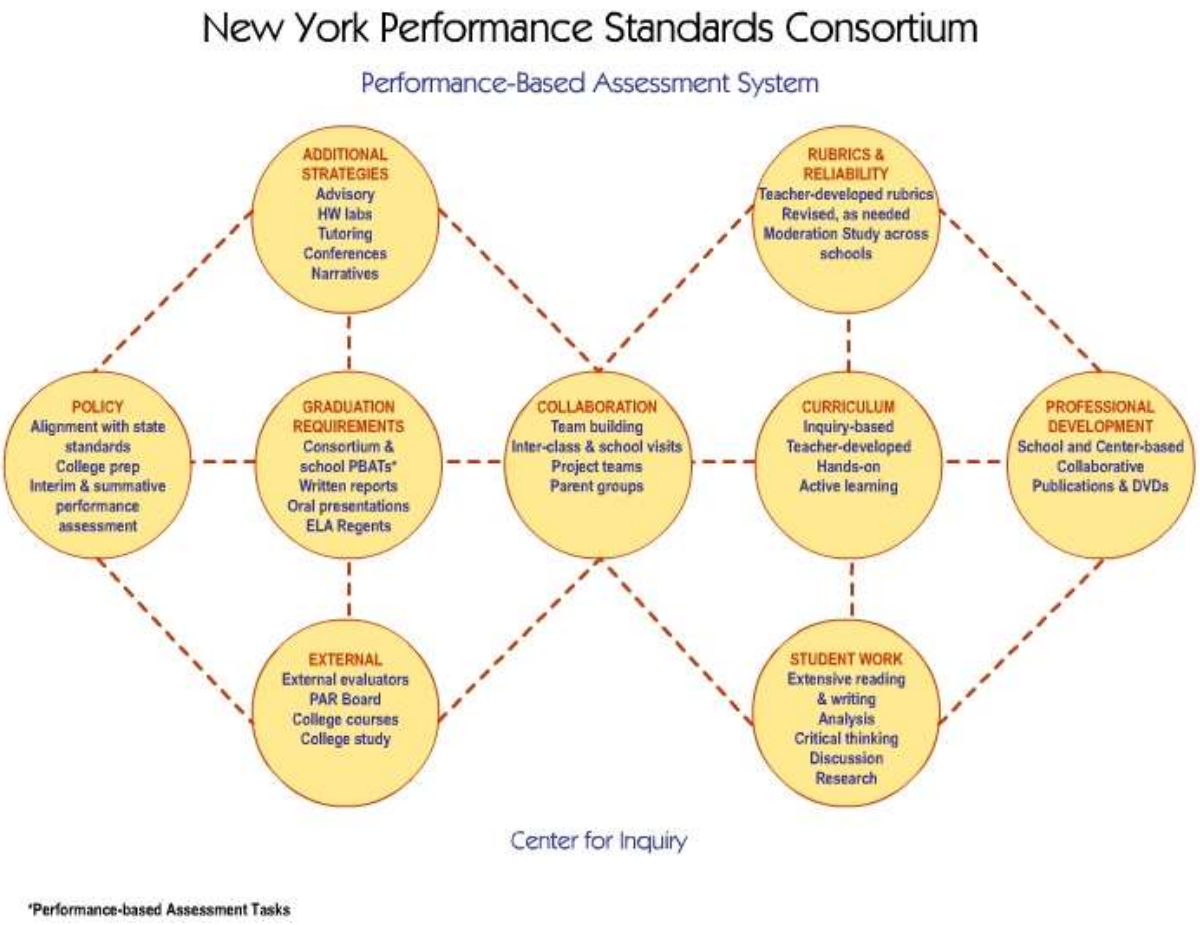
(III) Multiple measures must allow multiple opportunities to demonstrate achievement, be accessible to students at varying levels of proficiency, and utilize different methods for demonstrating achievement.

(IV) Assessments used shall meet appropriate technical standards to ensure the validity of the inferences likely to be drawn from the assessment results.

(V) While any one assessment may incorporate different methods (e.g., an exam with multiple-choice and extended response questions), multiple measures does not mean one assessment with several different components, nor only multiple opportunities to take the same assessment, nor two or more measures that are largely similar such as a state exam using mostly multiple-choice items and a state-mandated use of a norm-referenced test using similar item types or a district final or "benchmark" exam also using similar item types.

Note: This excerpt is at pages 6-7 of the FEA comprehensive recommendations, available on the web at http://www.edaccountability.org/NCLBlegrecs307.pdf.

B. New York Performance Standards Consortium Assessment System



http://performanceassessment.org/images/performance/NYPBAS_chart.jpg