

# 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. In response to concerns about No Child Left Behind's narrowing of curricula, some states have begun to use techniques they falsely label "multiple measures." Unfortunately, these are usually just multiple uses of the same statewide, standardized test results, not authentic multiple measures.

**Examples** of real multiple measures abound, including science labs or field work, from short tasks to extended projects; oral presentations in any subject; extended math problems that require application to real world uses; reading aloud and conversing with the teacher about a book; in-depth history reports, presented orally, in an essay, a PowerPoint, etc.; writing a paper in a second language; art or music projects; and answering questions from an expert panel about a project the student has done, much as doctoral candidates defend their theses. Documentation of teacher observations or interactions with the teacher can be useful, particularly with young children, if well structured. Many of these can be done individually or in groups (so long as the purpose is clear). This material can be organized so that it can be re-scored by other, independent educators, to ensure the accuracy of the classroom teacher, a process known as "moderation."

#### **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*: Dozens of NY high schools have a variance to use just one state exam (ELA) out of a mandated five. The Consortium uses the ELA test and four 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.

- *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.

- The *Coalition of Essential Schools* has documented a variety of ways in which member schools use performance assessments and multiple sources and types of evidence of student learning.

**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* uses multiple forms of assessment and relies on local assessments. Linda Darling-Hammond explains: “[In the 1970s,] 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.]

- *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.

- *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. (On formative assessment, see <http://www.fairtest.org/position-paper-assessment-learning>.)

- *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 (independent reviews of the assessment system) and moderation, both reliability and comparability can be established.

- The full paper is available at <http://www.fairtest.org/sites/default/files/MultipleMeasures.pdf>.