

Beyond Two Test Scores: Multiple Measures of Student Learning and School Accountability

Reasonable people understand a single test score does not define student learning. We must improve assessment systems and replace high-stakes, pass-fail testing policies with a system of multiple measures to evaluate student growth. The result: better teaching and learning that will benefit all students.

—NEA President Dennis Van Roekel

Students spend between six to eight hours in school per day, which amounts to approximately 1,080 to 1,440 hours per year. In that year, there are many opportunities for students to let their teachers and parents know what they have learned and for a school to let the community know about its progress. Students share their learned knowledge, for example, through their oral presentations and writings and on classroom and district-based assessments. However, the current version of the Elementary and Secondary Education Act (ESEA)—No Child

Left Behind (NCLB)—bases decisions on student and school success on a once-a-year set of standardized tests in only two subjects, reading and mathematics. Students have the right



to provide as many opportunities as possible to demonstrate what they have learned, and schools deserve more than just two test scores to show their progress.

One important alternative to the narrow testing that results from current federal law is the use of multiple measures of student learning. Multiple sources of evidence of student learning can include items such as teacher, district, or school assessments, assessments in subjects other than reading and mathematics, performance tasks, projects, and portfolios. Incorporating data from such a range of sources would end the reliance on a single test of student knowledge in two subjects to evaluate students for school accountability purposes. It could allow for a more reliable measure of student learning, encourage the teaching of a broader number of subjects and skills, and promote the teaching of higher-order thinking skills.¹

The case for multiple measures has strong roots in assessment research, including decades of skepticism over the use of single tests for high stakes decisions involving students. In a seminal 1999 statement on assessment by the American Educational Research Association (AERA),

American Psychological Association (APA), and National Council on Measurement in Education (NCME), *Standards for Educational and Psychological Testing*, the groups concluded², as summarized by the APA, that “any decision about a student’s continued education, such as retention, tracking, or graduation, should not be based on the results of a single test, but should include other relevant and valid information.”³

Multiple Measures in Context— Assessment Principles

Multiple measures can provide a better way to evaluate a school’s progress than NCLB’s narrow focus standardized tests, but they should be used in the context of the following assessment principles that will help ensure that the overall accountability system is fair and meaningful.

First, assessments should be designed primarily to enhance learning, not simply to get data for a report. When assessments are limited to certain types of tasks such as multiple choice and short answer questions, educators are encouraged to focus on similar tasks in the classroom, even though such tasks do not promote deep understanding and high proficiency in skills. Assessment tasks should enable students to demonstrate high levels of learning in a variety of contexts, some reflecting real world application of skills.

Second, assessments should measure growth in student learning from one point in time to another as well as attainment of standards. Measuring the growth of individual students can provide more feedback to students, teachers, parents, and schools than simply comparing, for example, the tests scores of eighth graders one year, and the tests scores of eighth graders the next year.

Third, tests should be used primarily for their designed purpose, and should be valid and reliable if used in an accountability system. These tests should also account

for the fact that many factors influence a student's performance beyond the classroom and the school.

Fourth, one or two test results should never be the sole indicators of student growth and achievement. The purposes of schooling go far beyond teaching students reading and math. Schools should prepare students to be self-directed, continuing learners, participants in democratic communities, socially adept, and responsible. Schools may be able to produce acceptable student scores in reading and mathematics while ignoring the whole child.

And fifth, test taking should not overwhelm a student's classroom experience, a teacher's instruction, or a school system's resources. For example, general school accountability tests that cover grade spans (e.g., once in grades 4-6 instead of in every grade) would free up time for instruction, help lessen the burden that testing places on school time and resources and diminish NCLB's excessive emphasis on "teaching to the test."

Using Multiple Measures

Within the broad context of fair and meaningful accountability systems, multiple measures of student learning can play an important role, particularly if they follow common protocols to allow comparisons at the school or district level.

But what do we mean exactly by "multiple measures"? According to researcher Susan Brookhart, one broad way to describe multiple measures is: (1) measures of different constructs; (2) different measures of the same construct; and (3) multiple opportunities to pass the same test. A construct, Brookhart explains, is the attribute being measured, "in education, often achievement in a specific domain."⁴

In the specific context of student learning, the 2010 paper, *Multiple Measures, A Definition and Examples from the U.S. and Other Nations*, offered the following summary:

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.⁵

The paper highlights several systems where multiple measures have been used:

- The Learning Record is a system developed for diverse learners in the areas of reading, writing, speaking, and listening, with a uniform structure for gathering and evaluating information. A student's progress (e.g. evidence showing understanding of reading and writing samples) is documented and placed numerically on a developmental scale. Moderation processes "have established adequate to superior inter-rater agreement between moderators and the teachers," thus supporting teacher judgment and allowing aggregation of classroom and school results. The system was growing in the U.S., but was "largely swept aside by NCLB requirements."
- The Work Sampling System for children age 3-8 has demonstrated "strong validity and reliability" and includes three elements: teacher observations using developmental guidelines and checklists "based on national content standards and current knowledge of child development"; portfolios of student work; and summary reports.
- The New York Performance Standards Consortium, a consortium of high schools, uses a "combination of consortium- and school-based performance assessments for both ESEA's Annual Yearly Progress (AYP) and for graduation requirements." The performance 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 scientific research, appreciation of and performance skills in the arts, service learning and school to career skills." Common rubrics are used in

scoring, and the Consortium “validated the use of the rubrics in four subjects through shared re-scoring of sample work.”

- The pre-NCLB Nebraska Statewide Teacher-led Assessment and Reporting System (STARS) included local assessments that met statewide standards or local equivalent standards and ensured consistent scoring. Local educators helped develop the assessments in many cases and they were reviewed by independent experts and periodically audited. Most of the assessments “incorporated more extended, classroom-based work, including tasks and projects.”
- Wyoming’s “Body of Evidence” system allows districts several options in developing a peer-reviewed system of assessments that ensures state graduation standards have been met.⁶

Another form of multiple measures is *through course* assessments, which consist of tests given at intervals throughout the year. These can be short answer items, performance tasks or projects that provide data that is summed together at the end of the year to provide a truly “summative” score.

Using computer technology also opens up new approaches that allow students to demonstrate their learning in increasingly efficient ways that provide multiple sources of achievement data. Students can respond online and have responses scored almost immediately. They can apply their knowledge and skills to computer-delivered scenarios and prompts not possible in paper and pencil contexts. One emerging approach to assessment involves the use of both paper and pencil and computer tasks. This approach reflects the growing use of blended instruction, computer and traditional media, and the contexts encountered in the world beyond the classroom.

Multiple Measures from a Variety of Subjects

In addition to the different types of data that can be gathered on reading and math, student learning indicators can include what they have learned in other subjects and how well they combine knowledge and skills across subjects. The Common Core State Standards for English Language Arts acknowledge this interdisciplinary characteristic of

education as being essential. It provides standards for English Language Arts that focus specifically on literacy as applied in science, social studies, and technology contexts. Multiple measures should include indicators of student achievement in other content areas beyond reading and mathematics.

Multiple Measures Promote a Complete Education

Within the broad context of accountability, multiple measures of student learning can play an important role in ensuring access to a complete education. When school success is based on data

related to learning a wide range of skills and knowledge in content beyond reading and math, students are likely to have access to a more complete curriculum.

Currently, the emphasis on reading and math test scores has led schools and districts to eliminate

health, physical education, and arts from the curriculum and curtail the time allotted for science and social studies. This is especially true in schools that serve students from low-income communities and lessens access to a complete education like that offered in more affluent schools.

High-achieving countries use multiple measures (multiple sources of evidence of varying types) to evaluate skills and knowledge needed for the demands of this dynamic, technological era.

— Linda Darling-Hammond

NEA’s Policy Recommendations

- A reauthorized ESEA should require that state plans include an assessment system that is accessible and balanced. The system should include innovative summative assessments based on a variety of ways for students to demonstrate achievement, as well as efficient, valid, and reliable data systems.
- The federal government should provide grants for the development of improved state and local assessment systems, and should allow states significant flexibility in the design and implementation of assessment systems.
- Assessment systems should be developed with the collaboration and agreement of educators and other

key stakeholders; should take into account the multiple factors impacting a student's learning beyond a teacher's control; and should be designed according to principles that allow their use with students of diverse abilities and diverse cultural and linguistic backgrounds.

- Education stakeholders should consider school assessment systems that involve multiple indicators and multiple measures of student learning over time, including measures that assess higher-order thinking skills and knowledge necessary for life in a global and interdependent, 21st century society.
- Multiple measures of student learning that should be considered include teacher-created assessments; district or school assessments; student work (papers, portfolios, projects, and presentations); teacher defined objectives for individual student growth; and high-quality, developmentally appropriate, standardized tests that provide valid, reliable, timely and meaningful information regarding student learning and growth.
- While state or local agencies may choose to administer their own assessments more frequently—particularly to help improve instruction in a timely manner—standardized tests mandated by the federal government should not occur more than once in each of three grade spans (e.g., 4-6, 7-9, 10-12) during a student's K-12 career.

RESOURCES

National Education Association (2006). *ESEA: It's Time for a Change! NEA's Positive Agenda for the ESEA Reauthorization*. <http://www.nea.org/assets/docs/HE/TM-NEAPositiveAgendafortheESEAREauthorization.pdf>.

NEA Policy Brief (2009): *Growth Models—An Update on the Effectiveness of Determining Student Progress and School Accountability*. <http://www.nea.org/assets/docs/HE/PB10aGrowthModels.pdf>.

National Center for Fair and Open Testing (Fairtest) (2010), *Multiple Measures: A Definition and Examples from the U.S. and Other Nations*. <http://www.fairtest.org/files/Multiple-Measures.pdf>.

Common Core Standards Initiative. <http://www.corestandards.org>.

Brookhart, S. (2009). *Accountability Policies and Measures: What We Know and What We Need*. NEA Research Department. <http://www.nea.org/assets/docs/HE/Accountability-PoliciesandMeasures.pdf>.

REFERENCES

- ¹ Multiple measure of student learning should be distinguished from multiple measures of school performance. The former involves using different ways to measure a student's progress. The latter refers to different ways to measure a school's progress (e.g. graduation rates, attendance, students taking challenging courses), one indicator of which can be valid measures of student learning. Multiple measures of student learning are also sometimes discussed in the area of teacher evaluation (see e.g. NEA's *New Policy Statement on Teacher Evaluation and Accountability* (2011) at <http://www.nea.org/grants/46326.htm>).
- ² American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*.
- ³ American Psychological Association (2001). *Appropriate Use of High-Stakes Testing in Our Nation's Schools*. Retrieved from <http://www.apa.org/pubs/info/brochures/testing.aspx>
- ⁴ Brookhart, S. (November 2009). *The Many Meanings of Multiple Measures*. ASCD: Education Leadership 67((3), 6-12. Retrieved from <http://www.ascd.org/publications/educational-leadership/nov09/vol67/num03/The-Many-Meanings-of-%C2%A3Multiple-Measures%C2%A3.aspx>. Brookhart describes the multiple ways that measures can be combined, including: "(1) conjunctive, in which the student or group must pass all measures; (2) compensatory, in which higher performance on one measure can compensate for lower performance on another; and (3) complementary, in which the student or group must achieve the standard on just one of the multiple measures." (Citing Chester, M. D. (2005). *Making valid and consistent inferences about school effectiveness from multiple measures*. Educational Measurement: Issues and Practice, 24(4), 40–52.)

⁵National Center for Fair and Open Testing (Fairtest) (2010), *Multiple Measures: A Definition and Examples from the U.S. and Other Nations*. Retrieved from <http://www.fairtest.org/files/MultipleMeasures.pdf>.

⁶For a detailed discussion of multiple measures in the context of high school graduation, see Darling-Hammond, L., Rustique-Forrester, E., & Pecheone, R. L. (2005). *Multiple measures approaches to high school graduation*. The School Redesign Network at Stanford University. Retrieved from www.srnleads.org/data/pdfs/multiple_measures.pdf.