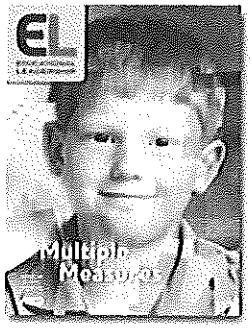
EDUCATIONAL LEADERSHIP



- Read Abstract

November 2009 | Volume **67** | Number **3**
Multiple Measures Pages 6-12

The Many Meanings of "Multiple Measures"

Susan M. Brookhart

To use multiple measures appropriately, start by understanding their purposes.

We wouldn't think of making most of our important life decisions on the basis of one measure alone. For example, people who are considering buying a house look at the house's age, condition, location, style, features, and construction, as well as the price of nearby homes. Doctors diagnosing an illness use multiple assessments: the patient's medical history, lab tests, answers to questions about how the patient feels, and so on. The question is, Why do education policymakers and practitioners sometimes opt to make important decisions based on only one indicator?

What Do We Mean by *Multiple Measures*?

Many people think of *multiple measures* in the plain English sense of the term, to mean using more than one score to make judgments about groups (such as classes, schools, and school districts) as well as individual students. The principle seems simple enough. As the National Council on Measurement in Education (1995) states in its *Code of Professional Responsibilities in Educational Measurement* (Section 6.7),

> *Persons who interpret, use, and communicate assessment results have a professional responsibility to use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions.*

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, Standard 13.7) confirms,

> *In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.*

In fact, Title I of the 1994 Improving America's Schools Act required the use of multiple measures to judge the performance of schools, and that language carried over unchanged, in 2001, to No Child Left Behind (NCLB):

> *Such assessments shall involve multiple up-to-date measures of student academic achievement, including*

measures that assess higher-order thinking skills and understanding.

Seems clear, but consider this. The 2004 NCLB guidelines count anything that measures higher-order thinking as "multiple measures." Some states count more than one opportunity to pass the same graduation test as "multiple measures." Are these interpretations what you had in mind when you read the three standards just listed?

To make good decisions about how to use multiple measures, both policymakers and practitioners need a clear understanding of what they mean by the term. Such an understanding begins with knowing what these measures are supposed to accomplish.

The Rationale for Multiple Measures

There are two important reasons to use multiple measures for decisions about education. The first is that multiple measures enhance *construct validity*. A *construct* is the attribute you are trying to measure. (In education, this is often achievement in a specific domain, but constructs also can be psychological traits, attitudes, and so on.) Construct validity is the degree to which any score conveys meaningful information about the attribute it measures.

We can't really get a full picture of Johnny's reading comprehension from one test score. The set of items or tasks on any one measure can't adequately represent the depth and breadth of complex concepts like reading comprehension or math problem-solving. Several measures, taken together, are likely to more adequately sample the things students should know and be able to do in the achievement domain being measured. Using more than one measure also helps us recognize performance variations caused by format, timing, and other logistical aspects of testing.

The second reason for using multiple measures is that they enhance *decision validity*. For any particular decision, there are usually several relevant types of information, each of which could have one or more measures. Johnny's reading comprehension is not the only important thing to know before we decide whether to place him in special education. We might also consider his achievement in other school subjects; his history with other interventions (for example, a reading support program); his affective responses to school; and his parents' observations of his work at home.

Another example: To decide whether a school is doing a good job, we need to consider several different achievement measures (reading, mathematics, and so on); as well as information about resources (personnel, financial, policy); processes (curriculum, instruction, school climate); and other school outcomes (safety, graduation rate, student and parent satisfaction).

In practice, there are many ways to define and apply the concept of multiple measures. Two questions are at issue. First, what counts as a "measure"? Second, how are the multiple measures combined? Here we'll discuss three different ways of counting what a measure is and three different ways of combining measures to make instructional decisions. If put these together, we end up with the nine different combinations shown in Figure 1 (p. 10).

Figure 1. Using Multiple Measures for Education Decisions

Three Ways to Combine Multiple Measures

Three Ways to Define Multiple Measures	**Conjunctive** **Student or school must pass all measures.**	**Compensatory** **Higher performance on one measure can compensate for lower performance on another.**	**Complementary** **Passing any one of several multiple measures suffices.**
Measures of different constructs	In Virginia, school accreditation ratings are based on students meeting achievement standards on tests in English, history/social science, mathematics, and science (with possible adjustments for ELL and transfer students and for preparing students for retakes of the state tests).	*U.S. News and World Report* compiles a list of "America's Best High Schools." One of its criteria involves computing a college-readiness index as a weighted average of advanced placement/International Baccalaureate participation rates and AP/IB performance quality.	The NCLB "safe harbor" provision means a school can meet its adequate yearly progress target if all subgroups meet the target percentage scoring proficient (achievement) or if the percentage of students who score below the proficient level in a subgroup decreases by 10 percent from the previous year (improvement).

Different measures of the same construct	An elementary school reading teacher requires a student to pass a reading comprehension test on at least two stories at the same reading level before allowing the student to read stories at the next higher reading level.	For a student's standards-based report card grade, under "Measures length to the nearest inch and/or centimeter," a teacher averages results from two quizzes and two performance assessments.	A teacher allows students to choose whether they will write a term paper or do a class presentation to show their understanding of Roosevelt's New Deal.
Multiple opportunities to pass the same test	In Louisiana, students who have met all graduation requirements except passing the graduate exit exam may continue to retake it—even after completing grade 12—until they pass.	A science teacher allows a student to retake a test that he or she failed after a unit on ecosystems and uses the average of the two test scores in the student's grade.	In Washington State, students in the class of 2013 will have to pass a mathematics test to graduate from high school. They may choose either the math portion of the state test or an Algebra I or Geometry end-of-course exam.

What Counts as a "Measure"?

"Multiple measures" describes at least three different ways of using more than one score: (1) measures of different constructs, (2) different measures of the same construct, and (3) multiple opportunities to pass the same test.

Measuring different constructs is helpful when we should base a decision on a combination of factors. For example, in the 1990s, as the standards movement was gaining momentum but before NCLB prescribed what sorts of measures states must report, states began experimenting with different indicator systems for school accountability. These systems included measures of school context (resources, student background variables, and so on); processes (curriculum coherence, leadership and teaching, and so on); and outcomes (student achievement, graduation rate, school safety, and so on). Indicator systems were designed because decisions about school effectiveness should be based on many different factors. Meaningful evaluations of outcomes, and especially decisions about what to change to bring about improvement, require that we also consider the context and process factors that work together to determine those outcomes.

Different measures of the same construct are especially helpful if the construct is some aspect of student achievement. To measure the construct thoroughly and to make sure all students have a chance to show what they know, several measures are better than one. If a student can't read and scores poorly on one assessment, additional measures of reading should confirm that fact; but if the student *can* read and performed poorly on one assessment for some other reason (perhaps an inability to connect with the stories or items on one particular test, or spatial difficulties that make it difficult to fill in bubble sheets efficiently), then additional measures will probably pick up his or her true capability.

Multiple opportunities to pass the same test may seem like an odd definition to include in the list. Nevertheless, in practice, this is sometimes called "multiple measures." For example, most states with graduation tests build in multiple opportunities for students to take the test.

How Are the Multiple Measures Combined?

Methods of combining information from multiple measures include (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 (Chester, 2005).

Most teachers' classroom grading policies are compensatory: They summarize students' scores on several achievement measures, usually either by calculating an average or by reviewing the whole set of measures with a rubric. Good performance on one measure can make up for poor performance on another. Typically, these classroom grades don't all measure the same construct—performance on a test and performance on a project tap different sets of knowledge and skills—but they are treated as though they do and summarized into a grade with one name ("Mathematics").

We can use multiple measures in a compensatory way at the school level, too. For example, Maryland's School Performance Index (SPI), which was used for state accountability before NCLB, judged a school's performance by combining 13 factors: percent satisfactory or better on each of six tested content areas for 3rd grade and for 5th grade, plus the school's overall attendance rate divided by the criterion of 94 (giving a school with 94 percent attendance a "perfect" score). Decisions about the schools' overall effectiveness were based on the total School Performance Index score. Other decisions—for example, those related to school improvement plans or curriculum—were based on the results of particular tests and subtests. Thus, the performance index not only provided an average measure of school quality, but also provided more fine-grained results to guide improvement (Schafer, 2003).

The current NCLB accountability system, at least as regards achievement, uses multiple measures in a conjunctive way. Districts must show adequate yearly progress overall, but also for every subgroup. Good results for one subgroup don't compensate for poor results for another subgroup. NCLB's safe harbor provision, however, uses complementary logic: A subgroup that does not achieve its annual performance goal can still "pass" if the percentage of students scoring below proficient in that subgroup decreases by 10 percent or more.

Multiple Measures Linked to Purpose

In many cases, individual educators don't have a choice about the application of multiple measures—for example, in their school's reporting of adequate yearly progress under NCLB. Even in these cases, though, it's important to have a clear understanding of what's going on. Knowing the nature of the measures and the combination method in any particular application of multiple measures helps us understand the results and the value of decisions or consequences based on those results.

Sometimes, however, educators do have a choice. The guiding principle for decisions about what measures to use and how to combine them should be *purpose*: What do you need to know, and why do you need to know it (Chester, 2005)? Here are some examples.

Classroom-Level Decisions

First, the bad example. Classroom rubrics sometimes mass multiple measures together in ways that distort the purpose of accurately reporting how well students have achieved learning goals. To evaluate student-created posters about different U.S. states, an elementary social studies teacher used a rubric consisting of four different measures: directions followed (5 points); information conveyed (10 points); creativity (10 points); and design/color/neatness (5 points). These are, arguably, measures of at least two different constructs: knowledge about states and poster-making skills. To arrive at the grade, the teacher added the points together (a compensatory method); only one-third of the resultant decision about achievement (10 points out of 30) was based on content. Thus, a grade intended to reflect achievement of a social studies objective actually largely reflected achievement of design and construction skills.

Now, a good example. A high school English teacher's class showed a wide range of ability to communicate in standard written English. The teacher wanted to assess the students' understanding of the plot of a novel the class had read. She used several different measures of the same construct (understanding of the plot) in a compensatory manner.

One measure was a test that had two parts: a written essay, which demonstrated students' ability to apply their understanding but was also, of course, affected by students' writing ability; and a multiple-choice section that didn't require writing but couldn't measure extended thinking about the novel. Another measure was an assignment in which students wrote open-ended questions at the end of each chapter; this task revealed students' thinking about the plot without requiring much formal writing. The teacher combined grades from all three measures to give a richer picture of plot understanding for all students in the class.

School-Level and Policy Decisions

If our main concern is to know for certain whether a school has reached a goal on a particular achievement construct (for example, a certain level of reading or mathematics performance), then we might want to use a compensatory approach combining multiple measures of that construct. If false negatives are a major concern—for example, if severe consequences are in place for failing to meet a standard—then we might want to use complementary multiple measures so that a school can pass by meeting the standard on any one measure. But if we are convinced that each of several measures is vital to quality, we'll probably want to use a conjunctive approach in which a school must pass all measures.

When states design high school graduation policies, a multiple-measures policy can help them avoid defining achievement narrowly as performance on one test. Darling-Hammond, Rustique-Forrester, and Pecheone (2005) reported that graduation rates stayed the same or declined slightly from 1998 to 2001 in five states that required students to pass an exit exam (Indiana, North Carolina, New York, Florida, and South Carolina). Four states that

used a multiple-measures approach to graduation during that time (New Jersey, Wisconsin, Pennsylvania, and Connecticut) fared better: Rates stayed the same in three states and rose in one. In addition, the graduation rates for these four states were higher overall in 2001 (73–86 percent) than those for the five exam-only states (51–67 percent).

Evaluations of school programs are also best accomplished by using multiple measures of different constructs. For example, suppose a district wanted to evaluate its K–12 science curriculum. One obvious measure would be performance on state exams mapped to district science standards—both absolute levels of achievement (status) and the amount of change in achievement (growth).

To make wise decisions about the science curriculum, however, the district would probably want to include other measures—for example, the number of graduates who go on to major in science or work in a scientific field, how students perceive the importance of science or how confident they feel as science learners, student participation in science-related clubs and activities, and so on. These are all different constructs, but they all have a bearing on making decisions about the science program.

Multiple Measures for Meaningful Decisions

The term *multiple measures* can mean many things. What's important is that multiple measures result in meaningful, useful decisions. A clear understanding of the many faces of multiple measures helps us think about the logic used in each case. Wise actions can only result if the measures and logic are right for their intended purposes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.

Chester, M. D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice, 24*(4), 40–52.

Darling-Hammond, L., Rustique-Forrester, E., & Pecheone, R. L. (2005). *Multiple measures approaches to high school graduation*. Stanford, CA: Stanford University School Redesign Network. Available: www.srnleads.org/data/pdfs/multiple_measures.pdf

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement*. [Online]. Available: www.natd.org/Code_of_Professional_Responsibilities.html

No Child Left Behind Act of 2001, Pub. L. No. 107-110, Sec. 1111(b)(3)(C)(vi). Available: www.ed.gov/policy/elsec/leg/esea02/pg2.html

Schafer, W. D. (2003). A state perspective on multiple measures and school accountability. *Educational Measurement: Issues and Practice, 22*(3), 27–31.

Susan M. Brookhart is Senior Research Associate at the Center for Advancing the Study of Teaching and Learning (CASTL), Duquesne University, Pittsburgh, Pennsylvania. She is the author of *Exploring Formative Assessment* (ASCD, 2009) and *How to Give Effective Feedback to Your Students* (ASCD, 2008); susanbrookhart@bresnan.net.

KEYWORDS

Click on keywords to see similar products:
assessment, accountability