Some Considerations of Multiple Measures in Assessment and School Accountability

Brian Gong and Richard Hill

The Center for Assessment

Presentation at the

Seminar on Using Multiple Measures and Indicators to Judge Schools' Adequate Yearly Progress under Title 1

Sponsored by CCSSO & US DOE

Washington, DC, March 23-24, 2001

Professional standards, legislative documents, and educational reformers call for "multiple measures"

- *Standards for Educational and Psychological Testing* (1999)
- Title 1 of ESEA
- AERA statement on high stakes testing
- NRC book on high stakes testing

- "Authentic assessment"/ "performance assessment"
- National science education standards
- "Multiple intelligences"

The term “multiple measures” has been used in many ways

1. test more than one content area (e.g., reading and mathematics);
2. assess mathematics using a combination of multiple choice and constructed response format test items;
3. assess writing using an on-demand test and a classroom-based portfolio;
4. assess school performance using a combination of academic tests and other indicators;
5. make progressively “higher stakes” decisions about schools using a combination of accountability scores and other reviews;

- have specified a process whereby a student may be promoted or graduate, upon demonstration of meeting certain criteria, even if the student does not pass the state’s on-demand test;

More examples of “multiple measures”?

- have specified several areas or criteria, all of which must be met in order for a school or student to be recognized for rewards/sanctions, an endorsed high school diploma, or other accountability consequence;
- have several assessment instruments that can be used by students of various proficiency or presentation/response needs;
- allow each student several opportunities to retake the test to determine whether s/he meets the minimum score to be promoted/graduate.
- double score every constructed response item on the tests used for high stakes student accountability;
- assess school performance using an average of at least two years’ of data;
- assess school performance using as many grades of students as practical.

Validity and reliability emphasize different aspects of multiple measures

- Validity - multiple measures
  - for adequate construct representation and
  - to allow demonstration of competence through a variety of presentation/response formats, modalities, and administration conditions
- Reliability - multiple measurements
  - to reduce to an acceptable level--or at least identify--the uncertainty of interpretation or probability of an undesirable consequence due to error
    - usually through use of more reliable instruments and/or procedures, and/or
    - through considering additional (repeated) data

Challenges of multiple measures

- Identifying suitable measures
- Collecting adequate measurements
- Combining data into a score or decision in a way that is:
  - useful
  - efficient
  - defensible

Consideration of goals and school accountability

Three different goals and associated models of accountability:

- "How has this school done in relation to the state standards or goals?" [status model, e.g., percent of students meeting or exceeding the standard]
- "Did this school improve enough to be on track to meet the state goals within the prescribed timeline?" [cohort/successive groups model, e.g., are fourth grade scores higher than before]
- "How much has this school's faculty helped their students improve from where the students were?" [student longitudinal model, e.g., compare same students' growth from end of grade four to end of grade five

Evaluating accountability models

For evaluating the Status, Successive Groups, and (Quasi) Longitudinal models:

- purpose and uses (e.g., stakes, reporting)
- policy considerations
- validity and reliability trade-offs
- operational requirements and costs/benefits
  - What is the right amount of testing time?
  - What quality assessments can you afford?
  - What infrastructure is needed (e.g., Dept. staffing, student database)

Fitting the goal to what is measured and calculated

Grade	Year 1	Year 2
3	a	b
4	c	d
5	e	f

Model	What is Calculated (Yr 2)
Status	b + d + f
Successive Groups	(b-a) + (d-c) + (f-e)
(Quasi) Longitudinal	(d-a) + (f-c)

Adequate Yearly Progress and models

Model	Growth Target	Adequate Criterion	Variants
Status	greater than previous year	"just measurable difference"	statistically significant
Successive Group	long-range goal	enough in one year to be on track to make long-range goal	comparison bands; all subgroups; regression-based expected growth
(Quasi) Longitudinal	year's growth		

Judging growth or progress

- Individual student performance --> student score (Year1, Class1, Student1, Content1)
- Aggregate scores (across content areas, students, classes, subgroups, and/or years)
- Generate growth targets and criteria
- Compare scores (to each other or to “growth target”)
- Make judgment about whether school made enough growth (considering target and error) using decision rules (single arithmetic; multiple conjunctive/compensatory; profile/holistic)

Some design issues that affect reliability and validity

- What is the biggest factor affecting the reliability of school accountability decisions?
- How can highly reliable assessments yield accountability decisions with low reliabilities (and vice versa)?
- How does retesting affect the reliability of conjunctive and compensatory decision rules?
- What is the difference between an effective weight an a nominal weight, and how do low effective weights affect reliability and validity?

Some design issues that affect reliability and validity - continued

- How can standards be set on multiple measures?
- Why is who is included essential for validity of accountability interpretations?
- What are “perverse incentives,” and how can accountability systems be designed to minimize them?
- How can local assessments be used in a comprehensive assessment system?

Pressures against multiple measures

- limited resources of money, available expertise, and time (to design, develop, administer, score; take away from instructional time)
- cost/benefit decisions favoring simpler systems for political and operational reasons; plan for incremental implementation;
- technical challenges, including developing, equating, and combining results from complex tasks;
- requests to “narrow” or focus the standards to promote potential for greater student or school success;
- political distrust or impatience with complex system that is seen as not dealing with persistent failure;
- little understanding by policy makers and state department staff of technical rationales underlying multiple measures/measurements

Multiple measures may provide specific benefits to validity and/or reliability

- increase the match of the assessment system with content and performance standards
- increase the validity of student-level and school-level results
- increase the reliability and validity of student-level and school-level results
- increase the fairness of assessment results
- increase the likelihood that schools will provide instruction in critical content areas and provide instruction in a variety of appropriate ways that emphasize skills reflected in content and performance standards (CCSSO, Winter, 2000)

Some interesting examples of multiple measures in state accountability systems

- Kentucky writing portfolios;
- Wyoming “Body of Evidence” system;
- Massachusetts multiple reviews for school assistance and sanction;
- Louisiana student promotion policies

Some other topics

- Technical issues in combining specific assessments and indicators
  - use of a common scale and/or index
    - e.g., combine test scores NRT and CRT; state and local (NRT, CRT, other); combine test scores and other indicators
  - combining non-standard and/or different combinations of measures
- “Sufficient” data for characterizing school(s)
- “Valid” interpretations and uses of accountability for effective and fair school improvement and services to individual students (disaggregation of data)