Multiple Measures: Toward Tiered Systems

CSE Report 607

Eva L. Baker
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

September 2003

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.2: Systems Design and Improvement: Ideal and Practical Models for Accountability and Assessment
Project Director: Eva L. Baker, CRESST/UCLA

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

MULTIPLE MEASURES: TOWARD TIERED SYSTEMS

Eva L. Baker

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

Abstract

This paper examines multiple measures of performance in school accountability systems from two perspectives: laterally (different indicators of different domains) and vertically (indicators that are at different levels of depth of the same domain). From these perspectives, organizational responsibility and instructional sensitivity are examined. In particular, alternative procedures are explored for integrating into the multiple measures concept external, uniform top-down measures and responsive, locally adaptive bottom-up measures.

The emphasis on testing as an instrument of reform and accountability has systematically increased worldwide over the last two decades, with the latest acceleration in the U.S. stimulated by federal legislation. Part of the policy evolution has been the recommendation for the use of multiple measures (MM) in any assessment or accountability system. Advocating the use of MM to make decisions has a long technical history, principally related to the reliability of measures and the goal of avoiding the undue influence of measurement error in making decisions about student status (*Standards for Educational and Psychological Testing*, American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 146; Millman, 1997). This concept of MM has been in the national policy rhetoric for a decade, dating at least from the discussions of the National Council on Education Standards and Testing (NCEST) in 1991-92. The MM precept has subsequently found its way into successive revisions of the Elementary and Secondary Education Act ([ESEA]; see Improving America's Schools Act [IASA], 1994; No Child Left Behind [NCLB], 2002) and has been intended to be part of the subsequent accountability systems (see the analyses by Henderson-Montero, Julian, & Yen, 2003). Because there are continuing concerns about over-testing students, that is, shifting time from instruction to make room for the growing number of required tests, it is important to clarify the range of reasons

supporting the use of more than one measure of performance. Justifications for MM have moved beyond the measurement error logic. Multiple measures can be seen as a way to assure fuller representation of a construct or as an operational way to influence what teachers teach. Equity interpretations suggest that MM may more adequately represent the possible ways teachers already interpret content standards and may legitimate and accommodate these different interpretations by teachers. Others see MM as a method to provide individual students more fair opportunities to display competency, demonstrating performance either using different methods or under different conditions, such as time for testing and potential for student revision of performance.

Such a cacophony of interpretation is an acknowledged part of the political process (Wildavsky, 1993, p. 117); that is, while MM are generally regarded as a desired practice, rationales on behalf of the MM are offered with markedly varying and somewhat contradictory arguments. Because the use of MM adds to the complexity, cost, and interpretability of performance, participants in education need to understand the range of potential consequences of the MM strategy in order to make smart choices about which types of MM get used and at what junctures. Let us start over and consider why might MM be worth the cost and energy they entail. An excellent analysis of MM use has been prepared by Gong and Hill (2001), in which they consider options, barriers, and uses. The analysis herein will consider six rationales for MM and conclude with some comments about accountability and educational quality.

Using Different Measures for Different Testing Purposes

One view of MM, departing from the error-avoidance approach, is that we need different measures to serve different assessment purposes. Tests intended to serve diagnostic functions for individuals may be developed and reported in very different ways from tests designed to monitor a system's progress toward reform. Both may well be needed to improve system or individual performance. Clearly the march of policy has been in the direction of accreting purposes to existing measures (see, for example, the regulations for NCLB, www.ed.gov; R. C. Atkinson, personal communication, 2001), and depending on the design vision that guided their development, some tests bear the weight better than others. Few, if any, have developed validity evidence in support of their expanded uses. Nonetheless, using MM to support different test purposes makes sense. One obvious difficulty is that this approach appears to be uneconomical, requiring cost of development and

recurring time of students and teachers. Another difficulty, even more challenging, is the manner by which a tight conceptual relationship is assured among different tests that are intended to measure the same general construct but for different purposes.

A design approach for solving this problem has been advanced by Baker and colleagues (Baker, 1997; Baker, Abedi, Linn, & Niemi, 1996; Niemi, 1996), which suggests that properly designed tests can generate useful information for different purposes by using different reporting strategies. Thus, system-oriented measures can be turned to instructional improvement purposes, or at least a beginning utility in student diagnosis, if assessment design criteria are developed and items are created that allow reasonably reliable reporting of details and reasons for errors. The assessment framework is designed to be valid at multiple levels, and different purposes are operationalized by different levels of summary. A second approach, illustrated convincingly by Black and Wiliam (1998) and a central conclusion offered by Pellegrino, Chudowsky, and Glaser (2001, p. 283), focuses on the upward aggregation of carefully designed classroom-focused examinations, where the purposes of measures are first to help children learn, and additional policy uses require different levels of summary of results. Error is presumably mitigated by the teachers' application of prior knowledge of children's performance. Extensive theoretical and practical advances are needed to make this approach a reality, including the application of technology and the creation of new capacity. Nonetheless, the use of assessment to improve learning has generated considerable interest. The model we have investigated is intended to guide both standardized and teacher-made tests. Either approach, if efficiently designed, would reduce the time mandated for external testing and thereby create an opportunity for the measurement of a broader band of subject matter expertise. Both approaches rely on high-quality design, on the use of learning as a basis for assessment, on conceptual clarity about the standards or constructs to be measured, and on a level of expertise in assessment by classroom teachers that likely well exceeds current levels of practice.

Using Different Measures to Broaden and Deepen Inferences About Learning

When the MM policy argument was put into print (*Raising Standards for American Education*, National Council on Education Standards and Testing, 1992, p. 12), the argument in its favor flowed from far different sources than those above. Of high interest during the late 1980s and early 1990s was the way in which

standardized and classroom tests might reflect both deeper and more intensive performance of students. Validity criteria for such performance-based assessments were promulgated (Linn, Baker, & Dunbar, 1991), and they emphasized the importance of using measures that mapped more strongly to cognitive views of learning, where the learner activated memory and applied and internalized knowledge as shaped by given contexts or problems. Arrows were shot at broad and shallow subject matter sampling, brief items, and multiple-choice response modes. Pulsing hearts, mine included, were aroused by the idea of extended tasks, real engagement in the cognition and knowledge demands, and the heightening of realistic application contexts for testing. The approaches advocated emphasized assessment methods that required students to demonstrate competence in tasks that approximated the way in which students would be expected to function outside of school. As a result, assessments were often to be integrated into instruction over a longer period of time (with feedback and reflection for student and teacher) and were intended to document the development of skill levels over a period of many days, or even a year or more.

Both the policy and technical communities had advocates for and antagonists against this new type of assessment, and the MM argument was mounted in the service of contradictory goals. As a result, MM was supported by coalitions to further conflicting goals. One set of proponents wanted MM provisions in legislation and policy guidance (following the IASA 1994 amendments) in an effort to protect the long-standing use of standardized achievement tests against the onslaught of performance assessment enthusiasm. The opposite side supported MM because of their beliefs that students needed a range of methods to display their knowledge and skills and that performance measures would broaden the quality of delivered educational services (see, for example, Koretz, 2003). Thus, as a compromise under either analysis, a given standard intended for achievement in a particular grade range might be measured both by a standardized achievement test and by a performance measure.

Deepening and intensifying assessment was a mission that carried certain, but unresolved, burdens that have undermined the validity argument for MM. One set involved the logistics of testing and included providing appropriate materials, standardization, and motivating teachers to do such assessments appropriately. Recurring costs of test design and scoring were at unprecedented levels. Some users of performance-focused tests decided that they needed to add normed tests for

public credibility (e.g., Kentucky). Weighting of tests has been a continuing technical issue (Chester, 2003; Schafer, 2003). Feasibility and relatively untried (and expensive) item development techniques in a number of sites led the majority of policymakers in the various states to believe that this approach to the use of MM was simply not appropriate for large-scale use. Some testing programs continue to include performance assessments in large-scale assessment for evaluation or system monitoring purposes, but for the most part, the idea of testing and reporting results for individual students wiped out the matrix sampling approach used by states and districts (for example, consider the demise of the Maryland School Performance Assessment Program [MSPAP]). The residual use of performance measures now occurs principally in two places. Logic impels the continuation of writing assessment as a "performance" or at least as a constructed response test. In addition, the new requirement for English language proficiency tests in No Child Left Behind (NCLB, 2002), where reading, writing, listening, and speaking skills are expected, makes a new place for such measurement approaches in the future. There are states and school districts with continuing commitment to deeper assessment. Whether their commitment can survive financial exigencies remains to be seen. The addition of science content standards and assessment is sure to support the use of technologically driven performance measures. There are those who believe that performance assessment will rise again with the support of technology (Baker, 2002) to remedy the problems of design, administration, and scoring costs.

Minimalist Implementations of Multiple Measures

A second conception of MM involves the array of measures in a system rather than the number that any particular student is asked to take. This interpretation can be observed in the use of assessments to meet program accountability requirements. During the negotiated rule-making activities in 1995, following the enactment of the Improving America's School Act (1994), a discussion was held to formalize the definition of MM in order to understand needed provisions of policy guidance. The negotiations involved conceiving of MM as variables other than achievement. These variables or indicators were termed "noncognitive" measures, a category that had been treated in the literature as affective and psychomotor performance (Bloom, 1956). The examples given in the discussion and surviving in policy guidance issued by the U.S. Department of Education included attendance, graduation rates, and dropout statistics. Many of these indicators suffered from a lack of standardization in their computation. Multiple measures in this instance meant summary indicators

of the system performance (including process measures) rather than measures of student academic accomplishments. During the same IASA discussion, the question was raised about whether a test battery, using the same format but addressing different content, would meet the MM expectation. Some thought so. Clearly, this evolution moved very far from the idea of using MM to reduce error around decisions about individual performance (reliability) or to increase the validity of measures. The degree to which introduction of these types of noncognitive measures reduces error in decisions affecting schools (i.e., which are designated in need of assistance) depends, of course, on the combination and weighting of such factors in either an index or a decision model.

Using Multiple Measures to Diversify Authority

A second use of MM for a system rather than for individuals occurs when states wish to encourage local districts to develop their own tests for classroom use. The local development could involve all grade levels to be tested, or only those "new" levels (probably four, in most settings) that are additional NCLB requirements. During the Congressional negotiations around NCLB, some states with a tradition of low centralization for education were looking for strategies that would allow them to use multiple local measures in their systems. In Maine and Nebraska, for example, there is an interest in using locally developed examinations to monitor state standards (http://www.state.me.us/education/lsalt/compassess.htm; http://www.nde.state.ne.us/starsdocs.html). The utility of this approach might well depend on the overlap among local assessments and the degree to which policymakers have faith in judgmental approaches to alignment. While such approaches garner skepticism from the measurement community, they are reasonably well accepted when the unit to which MM are applied is the state rather than the district. After all, since 1992, it has been agreed that different states may use different measures to implement ESEA requirements. Still, the question of technical capacity at a local level must be raised.

In the NCLB plan, a role for the National Assessment of Educational Progress (NAEP) was envisioned. The early intention was that NAEP would be used as a way to verify the gains shown on assessments used by the different states. This use is intended to serve as an external criterion. In the same vein, state assessments could be used as a potential external validity criterion for local measures. In the NCLB regulations (www.ed.gov), there are explicit criteria to be used if local assessments are to be employed. In either case, an argument can be made that (making the large

assumption that all measures have comparable quality) local or state measures ought to show considerably more growth than more distal measures simply because they are presumably more systematically connected to curriculum, instruction, and teacher professional development. However, the argument is probably swamped by differences in quality of the tests used in such a system.

Accommodations and Alternatives

A third example of MM in a system context is the use of accommodations and specific alternative measurement systems. Whether accommodated tests meet the standard of an "additional" system test is probably wide open to discussion and will be greatly determined by the degree of accommodations actually provided. It is clear, however, that some accommodated tests are substantial redesigns of the original measures intended for large-scale use. For example, tests in translation may not measure constructs similar to those intended for the majority population, and validity studies are usually not frequently conducted. This problem is complicated when, as in increasing communities, there is a large number of different primary languages spoken by a student group. Validity questions are often not well studied for accommodations made for students with disabilities of various sorts.

One may also very well raise questions about whether another type of assessment may be legitimated for those students who receive their education in nonstandard settings, such as those in juvenile facilities or in very small schools whose numbers do not reach the thresholds needed for inclusion in a state system. Most states have strategies that permit such alternative measures, but because of the transient nature of the population tested or the small numbers, it is difficult to establish the relationship of such performance to the expectations for the rest of the student population.

Multiple measures can also be used in an approach that blends some of the system and individual MM approaches described above. Districts or even schools may develop standards-based measures that supplement or expand the measurement base of statewide tests. Examples include districts in Chicago, Inglewood (CA), Los Angeles, and New York where classroom measures are used to assess the standards. The claim is that such measures provide another basis for student diagnosis and instructional improvement and that the artifacts of the assessments clearly communicate expectations to teachers and help them learn new ways of evaluating students. Other settings provide alternative tests for students

who have been unsuccessful in their performance on the major test—for instance, allowing students to take a project-based test if their performance on a standardized achievement test was below that needed for passing.

In summary, these uses of assessment may be characterized in different ways depending on whether we are focusing on the technical quality of the achievement measures or looking at a set of measures and indicators across an entire system. A compensatory justification at an individual level would be to provide additional opportunities (occasions, and/or methods) to assess performance. Compensatory justification at a system level would entail using local assessments to address standards in place of a state test. An alignment justification (see Commission on Instructionally Supportive Assessment, 2001) would be that local and classroom tests need to measure the key standards identified by the state that are otherwise impractical to assess on a large-scale basis. Another function, cycle completion, is to use local (classroom) assessments to fill in for the grades not included in the state test. Prior to NCLB, many districts tested students in these "off-years" by contracting for additional tests. Each of these characterizations is intended to provide a clearer picture of why MM are used, but when rationales are not explicit, the fragile goal of coherence is threatened. We will need operational definitions of coherence and documentation of the benefits of various approaches, or else we will shrink to one easy-to-administer, low-cost measure.

Using Multiple Measures to Broaden Our Definition of Educational Quality

Despite MM's paternity in reliability arguments, the mother of MM is validity and should exert full sway on the design and continuing evaluation of assessment and accountability systems (Baker, Linn, Herman, & Koretz, 2002). There is a series of deep questions that need to be answered about MM, such as: Are the current conceptions and operations of MM driving the system to be effective and efficient at the classroom level? This question generates a multitude of others: Is there a way for local assessments to matter? What do we actually mean when we say state assessments are aligned with standards, with classroom assessments, and with instruction? Are there degrees or elements of alignment that should count more or mitigate more? How should weighting of different assessments occur? How much redundancy do we want, and how much, if any, can we afford? How do we assure that the system promotes a broad conception of educational quality? Have we really bought into an educational culture that values only what is formally measured?

Limiting our expectations to those things that can be practically, economically (and at least validly) measured by large-scale tests will undoubtedly restrict experiences and explorations. We need to find and legitimate approaches that at once work to expand the range of educational goals, measure them in depth, and yield information that can be instructionally approached, and then weave the pieces into a strong and sensible fabric. Raising and resolving questions should strengthen, rather than weaken, the use of information in the service of learning.

Even though much of the discussion about assessment has focused on important concerns—fairly clear definitions of learning, understanding of local capacity, and concern for impact on various learners—we have leaned too much on simple devices—the format (multiple-choice or open-ended) or category (norm-referenced or criterion-referenced) of the tests we use. We may have lost heart in the pursuit of validity of the inferences or at least subjugated the meaning of the measures to logistics. In order to design a smart future for MM, we have to attend to the pervasive effect of accountability and study its impact on our understanding of reform as well as of the measurement of reform.

Moving education from an enterprise exclusively focused on process to an outcome emphasis is a relatively recent policy goal, and missteps and false starts are to be expected. The only way we can see whether we are making progress is to be mindful of the goals and consequences of our MM actions.

References

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

Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36*, 247-254.

Baker, E. L. (2002). Design of automated authoring systems for tests. In National Research Council, Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education (Eds.), *Technology and assessment: Thinking ahead: Proceedings from a workshop* (pp. 79-89). Washington, DC: National Academy Press.

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research, 89*, 197-205.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002, Winter). From the directors: Standards for educational accountability systems. *The CRESST Line*, pp. 1-4.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of education goals. Handbook 1: Cognitive domain*. New York: Longmans, Green & Company.

Chester, M. D. (2003). Multiple measures and promotion decisions: The way the measures are combined is almost as important as the measures. *Educational Measurement: Issues & Practice, 22*(2), 32-41.

Commission on Instructionally Supportive Assessment. (2001). *Building tests to support instruction and accountability: A guide for policymakers* (James Popham, Commission Chair). Washington, DC: Author. Retrieved June 29, 2002, from http://www.aasa.org.

Gong, B., & Hill, R. (2001, March). *Some considerations of multiple measures in assessment and school accountability*. Presentation at the Seminar on Using Multiple Measures and Indicators to Judge Schools' Adequate Yearly Progress Under Title I (Sponsored by CCSSO & U.S. DOE), Washington, DC.

Henderson-Montero, D. L., Julian, M. W., & Yen, W. M. (2003). Multiple measures: Examination of alternative design and analysis models. *Educational Measurement: Issues & Practice, 22*(2), 7-12.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues & Practice, 22*(2), 18-26.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21. (ERIC Document Reproduction Service No. EJ436999)

Millman, J. (Ed.) (1997). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Press.

National Council on Education Standards and Testing. (1992). *Raising standards for American education. A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people.* Washington, DC: U.S. Government Printing Office.

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representation, problem solutions, justifications, and explanations. *Journal of Educational Research, 89,* 351-363.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessments* (Committee on the Foundations of Assessment; Board on Testing and Assessment, Center for Education. Division on Behavioral and Social Sciences and Education, National Research Council). Washington, DC: National Academy Press.

Schafer, W. D. (2003). A state perspective on multiple measures in school accountability. *Educational Measurement: Issues & Practice, 22*(2), 27-31.

Wildavsky, A. (1993). *Speaking truth to power: The art and craft of policy analysis.* New Brunswick, NJ: Transaction Publishers.