The Ideal Role of Large-Scale Testing in a Comprehensive Assessment System

Charles A. DePascale
National Center for the Improvement of Educational Assessment

Abstract

The role of large-scale assessment in public education has grown tremendously since the mid-1980s and unquestionably will continue to grow with the implementation of the assessment and accountability requirements of the No Child Left Behind Act. In the rush to meet the demand to measure validly and reliably the performance of all students, however, it must not be forgotten that large-scale assessment is only one component of a comprehensive assessment system. The factors that led to the predominance of large-scale assessment are reviewed and the appropriate role of large-scale assessment in a comprehensive assessment system is discussed.

Introduction

It is not an exaggeration to assert that large-scale assessment is a greater force in K-12 public education in the United States at this time than it has been at any other time in U.S. history. Since 1990, large-scale assessment at the national and state levels has been characterized by increases in the number of tests administered, more demanding content and performance standards, and higher stakes for students and schools. At the national level, the National Assessment of Educational Progress (NAEP) has shifted from the 'judgment-free' reporting of long-term trends in national and regional performance (i.e., what students know and can do) to the reporting of state-by-state results in terms of judgment laden performance standards (i.e., what students should be able to do). At the state level, whether in response to the requirements of the Improving America's Schools Act of 1994, state education reform efforts, or the impact of the standards-based movement, in general, mandated state assessments have increased in quantity, content and performance demands, as well as stakes for students and schools. In addition, requirements of the Individuals with Disabilities Education Act of 1997 (IDEA97) have broadened the pool of students participating in large-scale assessments.

Furthermore, the role of large-scale assessment is likely to increase in the next several years. The federal No Child Left Behind Act of 2001 (NCLB) requires states to administer annual assessments to all students in reading and mathematics at grades 3 through 8 and once at grades 10-12 by the 2005-2006 school year, and also requires the annual assessment of the English proficiency of English language learners. Increased testing at the state level will be accompanied by more frequent state NAEP administrations in reading and mathematics.

Associated with the increased quantity and stakes of large-scale assessments are increased demands for information from these assessments. On one level, the demand is for a more rapid delivery of results from the assessments. On another level, the demand is for more in-depth reporting of assessment results at the school and student level. NCLB requires states to provide interpretive, descriptive, and diagnostic reports at the student level. A report by the Commission on Instructionally Supportive Assessment (Popham, 2002) calls for standard-by-standard reporting of student results on state assessments to allow "educators to evaluate the effectiveness of their instruction related to each standard and to improve that instruction when warranted" (p 14).

To meet the demands of NCLB, testing companies, states, and the federal government are focused primarily on making current large-scale assessments more efficient and effective. The goal is to produce more results that are more accurate more quickly so that those results can be more useful to more schools and more students. Of course, through the increased use of technology, it is expected that the goal should be accomplished with less testing time, less cost, less reliance on human scorers, and less standardization. In short, the goal is to develop methods to use large-scale assessment to better measure the achievement of all students.

Implicit in that goal is the assumption that large-scale assessments are an appropriate tool to measure the performance of all students. At the very least, there is the belief that the current model of large-scale assessment is necessary to ensure comparability of measurement across students, schools, and states. However, the current model of large-scale assessment use (i.e., annual assessment of all students in all content areas to produce individual student results) is not necessarily the only or the most appropriate long-term use of large-scale assessment or the best method to measure the academic achievement of all students.

In this paper, I attempt to provide a rationale for an alternative model for the use of large-scale assessment – a model that calls for a more limited (and well-defined) role for large-scale assessment in a comprehensive system of assessment at the local, state, and to some extent, national levels. The concept of a comprehensive assessment system is not novel, original, or innovative. In fact, with regard to the role of large-scale assessment, on the surface it may appear as simply a call for a return to an earlier time in which large-scale assessment served a different purpose in K-12 public education. The difference, however, is that in that earlier time large-scale assessment was often an isolated component in an incoherent collection of assessments.

This paper is divided into three major sections. The first section provides a definition of large-scale assessment and discusses the forces that led it to evolve to its current state. The second section discusses the role and appropriate use of large-scale assessments in public education. The third section describes a comprehensive assessment system and the role that large-scale assessment plays within such a system.

The Evolution of Large-Scale Assessment

What is large-scale assessment? The Montana Office of Public Instruction provides a succinct answer to the question in a Q&A document describing the Montana statewide norm-referenced testing program: "Large-scale assessment means tests are administered to large numbers of students, such as those in a district or state," (Montana Office of Public Instruction, 2001). For the purposes of this paper, the term large-scale assessment will refer primarily to tests administered as part of statewide assessment programs.

A natural follow-up to the initial question "What is large-scale assessment?" is the question: Why do we need large-scale assessment? Popham (2001) provides the following response in describing the "primary measurement mission" of large-scale assessment programs:

> It's all about *accountability*. Large-scale assessment programs, the bulk of which are of the high-stakes variety, are in place chiefly because someone believes that the annual collection of students' achievement scores will allow the public and educational policymakers (such as state school board members or state legislators) to see if educators are performing satisfactorily. (p. 34)

Popham (2001) and Landau, Vohs, and Romano (1999) both note that most statewide programs also purport to have an instructional component. That is, in some form, states claim that one purpose of the statewide assessment system is to improve instruction. In Massachusetts, this is reflected in the following statement describing the purpose of the Massachusetts Comprehensive Assessment System (MCAS):

> The primary goal of Education Reform is to improve student performance. MCAS serves two main purposes that focus on achieving that goal. First, it is designed to improve classroom instruction and assessment by: (a) providing specific feedback that can be used to improve the quality of school-wide, classroom, and even individualized student instructional programs; and (b) modeling effective assessment approaches that can be used in the classroom. Second, it serves as an accountability tool for measuring the performance of individual students, schools, and districts against established state standards. (Massachusetts Department of Education, 2002, p. 10)

A balance scale measuring the relative importance of accountability and instructional benefits in the decision to implement a large-scale assessment program would likely lean heavily toward accountability. The form and format of many large-scale assessments (i.e., including lengthy passages from a wide variety of sources and a mix of long- and short-answer constructed-response items), however, reflect the concern about the tests' instructional contribution and reveal a significant portion of the recent history of large-scale assessment.

The roots of the current period in large-scale assessment history are well documented and can be traced back to the mid-1980s and a series of loosely related events (Popham, 2001; Herman, 1997). It can be argued that the modern educational testing period began with the passage of the Elementary and Secondary Education Act of 1965 (ESEA), and by the mid-1980s a majority of states were administering some type of large-scale assessment program (Rothman, 1995). However, large-scale testing as we know it today began to take shape with the publication of *A Nation at Risk* in 1983. The report prepared for the United States Department of Education by the National Commission on Excellence in Education called for the adoption of rigorous and measurable standards along with higher expectations for students. As one step toward implementing that recommendation, the Commission suggested:

> Standardized tests of achievement (not to be confused with aptitude tests) should be administered at major transition points from one level of schooling to another and particularly from high school to college or work. The purposes of these tests would be to [determine]: (a) the student's credentials; (b) the need for remedial intervention; and (c) the opportunity for advanced or accelerated work. The tests should be administered as part of a nationwide (but not Federal) system of State and local standardized tests. This system should include other diagnostic procedures that assist teachers and students to evaluate student progress. (Recommendation B.3)

The findings and recommendations of the report fueled a growing sense of dissatisfaction with the performance of students in public schools and was a catalyst in a flurry of education reform efforts in states throughout the country. At roughly the same time, a 1987 report by a West Virginia physician, John Cannell, raised concerns about the appropriate use of traditional, norm-referenced, standardized tests. Cannell's findings that almost all school districts and a vast majority of students score above the national average on norm-referenced tests became widely known as the "Lake Wobegon Effect" (Cannell, 1987, Editorial Projects in Education, 1997)

The conclusions drawn from the Cannell study (in conjunction with the findings of subsequent studies on teachers and testing) heightened awareness that teachers "teach to the test." In one sense, "teaching to the test" can be interpreted as directly teaching the content of the items included on a standardized test administered repeatedly year after year. In another sense, "teaching to the test" refers to the influence that the content and format of state tests have on teachers and students. As Wiggins (1993) explains:

> Tests *teach*. Their form and their content teach the student what kinds of challenges adults (seem to) value. If we keep testing for what is easy and uncontroversial, as we now do, we will mislead the student as to what work is of most value. This is the shortcoming of almost all state testing programs. Generations of students and teachers alike have fixated on the kinds of questions asked on the tests – to the detriment of more complex and performance-based objectives that the state cannot possibly test en masse. (p. 42)

Why would state testing programs (i.e., standardized tests) exert such a strong influence on the teaching and testing behaviors of teachers – particularly at a time when the content of the tests was relatively simple and the stakes associated with those tests were relatively low? Part of the answer is that in the absence of statewide curriculum frameworks or standards, the content of the test becomes the de facto state curriculum. A second part of the answer lies in the deep void that existed in teachers' training in and understanding of assessment design and use (Gullickson, 1985; Wiggins, 1993).

The problem of a lack of statewide curriculum frameworks and standards lessened as states across the country began to develop and implement standards in response to state reform efforts, a growing standards movement, and finally, to meet the requirements of Improving America's Schools Act of 1994.

Unfortunately, there was no corresponding solution to the assessment side of the problem. Without an easy solution (or, in fact, any feasible solution) to improve teachers' assessment practices, the decision was made to change large-scale assessment to better model effective local assessment– if it is a given that teachers are going to teach to the test, give them a test worth teaching to. The call for an end to a reliance solely on multiple-choice tests and for more "authentic" assessment of student performance resulted in several changes in the content and format of large-scale assessment in the late 1980s and early 1990s such as

- Direct writing assessment in several states across the country,
- The use of constructed-response items requiring student responses longer than one or two sentences in states such as Kentucky and Maryland,
- Statewide portfolio assessment in states such as Vermont and Kentucky, and
- The birth of collaborative programs such as the New Standards Project and Council of Chief State School Officers State Collaborative on Assessment and Student Standards (CCSSO-SCASS) projects to develop new forms of assessments to measure high standards. (Rothman, 1995)

In summary, the emphasis on accountability and high standards combined with a lack of confidence in local educators' ability to assess students resulted in large-scale testing a) becoming the primary vehicle to assess all students, and b) serving as a model for local assessment. The current state of large-scale assessment in public education evolved in response to a perceived need. In large part, large-scale assessment expanded to fill the assessment and accountability void left by classroom and local assessment. This was not designed as a long-term solution to address the causes of the void – a task beyond the scope of the assessment community. Rather, the increased emphasis on large-scale assessment, at best, was a temporary patch or short-term solution. The long-term solution requires filling that void with valid and reliable results from local assessment systems at the district, school, and classroom levels.

The Role of Large-Scale Assessment

In the previous section, two primary roles of current large-scale assessment were discussed – modeling and accountability. Accountability can be subdivided into the distinct categories of school accountability and student accountability.

Is the current role of large-scale assessment appropriate?

In one respect, evaluating the appropriateness of the role of large-scale assessment in public education involves defining the role of the state in public education. The role of large-scale assessment should be consistent with the role of the state. Reaching consensus on what the role of the state should be in public education, however, is not a simple task. In New Hampshire, where the percentage of state aid to education is historically among the lowest in the nation, a major focus of a series of school funding lawsuits that consumed the state for a large portion of the 1990s (Claremont School District v. Governor, 1993, 1997) was defining the role of the state in providing an *adequate* education to all students.

In January 2000, the Massachusetts Board of Education identified two critical areas that defined the role of the state Board in accomplishing the overall goal of raising student achievement: accountability and creating effective schools. Within those two areas, however, there was a keen interest in finding the balance between state control and local control. Regarding creating effective schools, for example, there was general agreement that a major component was local autonomy and strong local leadership within the school and community. This is consistent with the literature on effective schools and is also consistent with the Massachusetts Education Reform Law of 1993 (Eiseman, 2003; Tappan, 2003). The Board expressed a similar need to define the limits of the state's role in the area of accountability.

The Massachusetts Board described a state role in public education that is based largely on communicating best practices, monitoring, and auditing (in addition to providing the resources for schools to succeed). Under such a system, the primary function of large-scale assessment is accountability – to provide the state with information on whether local districts and schools are meeting their achievement goals. Note that the emphasis here is on auditing the performance of groups of students within local districts and schools not on auditing the performance of individual students. Monitoring or auditing the performance of groups or subgroups of students is a conceptually distinct task from monitoring the performance of individual students.

Only when the role of the assessment is defined is it possible to direct attention to the design of the assessment – *form follows function*. A large-scale assessment designed for accountability may look quite different than an assessment designed to model effective classroom assessment and instruction. A large-scale assessment designed to monitor district and school performance may look quite different than an assessment designed to measure student performance. It is this type of distinction that former Assistant Secretary of Education Susan Neuman expressed when she called for states to consider a return to

tests based largely on multiple-choice items in her keynote address to the 2002 CCSSO National Conference on Large-Scale Assessment. Such an instrument would not be intended to be a *test that taught* to paraphrase Wiggins.

Of course, whether a state intends the test to teach and whether the test does teach are two separate issues. It is probably safe to conclude that it was not the intent of the states for multiple-choice tests to become the driving force in local instruction and assessment during the 1970s and 1980s. The key is to determine the factors that would change the dynamics between large-scale assessment and local assessment and instruction.

Movement toward a Comprehensive Assessment System

Carr and Harris (2001) describe a comprehensive assessment system that draws on data from the state/national level, the district/school level, and the classroom level to a) improve education, b) determine success, and c) provide feedback to relevant stakeholders (e.g., students, teachers, policy makers, the community). For large-scale assessment to assume the role of a tool used to monitor district and school performance it must be one component of a comprehensive assessment system. That is, there must be other components in place at the district, school, and classroom level to provide valid and reliable information about the performance of individual students.

As discussed previously, large-scale assessment assumed its current role to fill a void. The first component of the void was the absence of established curriculum frameworks or learning standards. The second component was limited knowledge or understanding of the principles and methods of assessment at the local level. One question that must be answered, therefore, is whether there has been any change in the underlying causes of the problem that led to large-scale assessment assuming its current role.

There appears to have been considerable progress in the area of states establishing curriculum frameworks and learning standards. Across the country, virtually all states have established standards that serve as an explicitly mandated curriculum, an implicitly mandated curriculum, or at least a model curriculum for all schools. Clarke et al. (2003) report on benefits of the development and implementation of state curriculum frameworks in Massachusetts:

- The linking of district- and school-level curricula to the state standards,
- The redefining of classroom work in response to the standards, and
- Curriculum spiraling – the vertical alignment of curricula across grade levels.

To the extent that a) school curricula are aligned with the state curriculum standards, and b) states' large-scale assessments are aligned with state curriculum standards (e.g., adequately sample the breadth and depth of those standards), the need and desire to teach to the state test will be diminished. At the very least, the difference between teaching to the standards and teaching to the test should become less distinct.

Unfortunately, there is little evidence of a sea change in classroom teachers' level of understanding of assessment and use of effective assessment practices in the classroom.

At an end-of-millennium symposium sponsored by the National Council on Measurement in Education (NCME), Richard Stiggins (2001) reported, "the state of classroom assessment affairs is dismal. It has been so for decades. As a result, harm has been and is being done to students, and I believe the time has come for that to end."

Across the country, however, there are some ripples of encouragement:

McMillan (2001), in a study of Virginia secondary teachers' classroom assessment practices, found the following in relation to teachers' assessment use:

- Essay-type questions are used only slightly less frequently than objective tests,
- There is considerable use of student projects and performance assessments,
- Assessments measure understanding the most, although there was also a strong emphasis on assessments that measure both reasoning and application,
- Assessments that measure recall were used the least, although they are still used quite a bit.
- There is greater reliance on teacher-developed instruments and very little reliance on assessments provided by publishers.

As McMillan noted, it is also worthwhile to note that this study was conducted in a state that, with the exception of a direct writing assessment, relies exclusively on multiple-choice tests for its statewide assessment.

At a workshop on bridging the gap between classroom and large-scale assessment sponsored by the National Research Council, Eva Baker presented a description of "model-based assessment" currently being tested in the Los Angeles Unified School District. Baker presented five reasons why some schools are successful in using assessment knowledge:

- A focus on learning (students and adults)
- Constant use of *appropriate* information (formal and informal)
- Focus on feedback and change
- Public display and exchange
- Community pride in outcomes of students and place. (Baker, 2003)

Baker also lists 'congruence or peace with external mandates' as a context for success of knowledge-based reform.

A session at the 2003 National Conference on Large-Scale Assessment, *Lessons from Nebraska, Maine, and Vermont: Building Local Assessment Capacity in School Districts,* highlighted the efforts of Nebraska, Maine, and Vermont to integrate local assessment into their statewide assessment programs to varying degrees. At the heart of each state's efforts was an acknowledgement of the need for continuing professional development and support to increase the assessment literacy of teachers, administrators, and the community.

Conclusion

Large-scale assessment is at a crossroad. With the focus on NCLB and its required large-scale state tests, the importance of large-scale assessment in K-12 public education has never been greater. With the advent of emerging technologies that will facilitate large-scale assessment administration, scoring, and the reporting of results, the danger is great that in the coming years we will travel so far down the large-scale assessment road that it will become impossible to strike the appropriate balance between classroom, local, and large-scale assessment.

Large-scale assessment will never occupy its proper role in a comprehensive assessment system until the other components of the system are established enough to provide credible assessment information that is largely consistent with the results of large-scale assessment. That is not to argue that classroom assessment results are somehow inferior to large-scale assessment results. Rather, the argument is that in a comprehensive system there is congruence among the results of the local and large-scale assessment. When classroom, local, and large-scale assessments are aligned with the same content and performance standards, and quality assessment exists at all levels, the influence of the large-scale tests will diminish.

As Stiggins (2001) notes, classroom assessment is still a long way from providing quality information and/or gaining the required credibility to assume its role in the comprehensive assessment system. However, the time is now to acknowledge that large-scale assessment cannot fulfill all of our assessment needs and to direct our efforts to striking the proper balance.

References

Baker, E.L. (2003). *Model-based Assessment: Why, what, how, how good, what next, and why not?* Presented at the National Research Council Bridging the Gap Between Classroom and Large-Scale Assessment Workshop. Washington, D.C. January 2003.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends of Education.

Carr, J.F. & Harris, D.E (2001). *Succeeding with Standards: Linking curriculum, assessment, and action planning*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Claremont School District v. Governor (1993). 138 N.H. 183, 635 A.2d 1375.

Claremont School District v. Governor (1997). 142 N.H. 462, 703 A.2d 1353.

Clarke, M., Shore, A., Rhoades, K, Abrams, L, Miao, J, & Li, J. (2003). *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from Interviews with Educators in Low-, Medium-, and High-Stakes State*. National Board on Educational Testing and Public Policy. Boston College. www.bc.edu/research/nbetpp/statements/nbr1.pdf.

Editorial Projects in Education (1997). *Quality Counts 1997: A report card on the condition of education in the 50 states*. West Virginia.

Eiseman, J.W. (2003, April). *Four successful comprehensive school reform scenarios*. Paper presented at the 35th Annual Meeting of the New England Educational Research Organization. Portsmouth, New Hampshire.

Gullickson, A.R. (1985). *Student evaluation techniques and their relationship to grade and curriculum*. Journal of Educational Research, 79 96-100.

Herman, J. (1997). *Large-Scale Assessment in Support of School Reform: Lessons in the Search for Alternative Measures*. CSE Technical Report 446. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles. cresst96.cse.ucla.edu/Reports/TECH446.pdf.

Landau, J.K., Vohs, J.R., & Romano, C.A. (1999) *Statewide Assessment: Policy Issues, Questions, and Strategies*. PEER Information Brief. The Federation for Children with Special Needs. www.fcsn.org/peer/ess/pdf/assessmentsw.pdf.

Massachusetts Department of Education (2002). *Spring 2002 MCAS Tests: Summary of State Results.* Malden, Massachusetts: Massachusetts Department of Education. www.doe.mass.edu/mcas/2002/results/summary.pdf.

McMillan, J.H. (2001, Spring). *Secondary Teachers' Classroom Assessment and Grading Practices*. Educational Measurement: Issues and Practice, 20-32.

Montana Office of Public Instruction (2001, August). *Release of IOWA test scores*, *MontCAS memo*, p. 5. www.opi.state.mt.us/PDF/Assessment/MontCas.pdf.

National Commission on Excellence in Education (1983, April). *A Nation At Risk: The imperative for educational reform*. A report to the Nation and the Secretary of Education, United States Department of Education. www.ed.gov/pubs/NatAtRisk.

Popham W.J. (2001). *The Truth About Testing: An educator's call to action*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Popham W.J. (2002). *Implementing ESEA's Testing Provisions: Guidance from an Independent Commission's Requirements.* The Commission on Instructionally Supportive Assessment. www.aasa.org/issues_and_insights/issues_dept/Commiss.n_Report_Book.pdf.

Rothman, R. (1995). *Measuring Up: Standards Assessment, and School Reform*. San Francisco: Jossey-Bass Publishers.

Stiggins, R.J. (2001, Fall). *The Unfulfilled Promise of Classroom Assessment*. Educational Measurement: Issues and Practice. 5-15.

Tappan R (2003). *Characteristics of Highly Improved Schools: A Case Study of Selected Schools in Economically Disadvantaged Districts*. Paper presented at the Reidy Interactive Lecture Series. Nashua, New Hampshire. www.nciea.org/publications/HighlyImpSchools_Tappan03.pdf.

Wiggins, G.P. (1993). *Assessing Student Performance: Exploring the Purpose and Limits of Testing*. San Francisco: Jossey-Bass Publishers.