# Putting Large-Scale Assessment In Perspective: The Ideal Role of Large-Scale Assessment in a Comprehensive Assessment System

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

## Introduction

It would not be 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, more demanding content and performance standards, and higher stakes for students and schools. At the national level, the National Assessment of Education 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, and stakes for students and schools. In addition, requirements of the Individuals with Disabilities Education Act of 1997 (IDEA97) has broadened the pool of students participating in large-scale assessments.

Furthermore, it would be an understatement to argue that 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 in reading and mathematics at grades 3 through 8 and once at grades 10-12 by the 2005-2006 school year. NCLB also requires an annual state assessment in science at the elementary, middle, and secondary levels by the 2006-2007 school year. In addition to those annual statewide assessments administered to all students, NCLB also requires prescribes the annual assessment of the English proficiency of English language learners. Finally, the increased testing at the state level will be accompanied by more frequent administration of the state NAEP administrations in reading and mathematics.

As one might expect, associated with the increased quantity and stakes of large-scale assessments are increased demands for information from large-scale assessments. On one level, the demand is for a more rapid delivery of results from the assessments. Beginning with the 2003 NAEP administration, the timetable for delivering NAEP results following each administration will be shortened by more than one year with NAEP results being released no more than six months following each administration. Under NCLB, results from state assessments must be returned quickly enough to allow sufficient time for districts and schools that fail to make adequate yearly progress or are declared in need of improvement to implement required programs and services prior to the beginning of the school year. On another level, the demand is for more in-depth

Paper presented at the 2003 Annual Meeting of the American Educational Research Association
Chicago, Illinois
April 2003

reporting of assessment results at the school and student level. For example, it is commonplace to hear discussion of the need for assessment results to provide diagnostic information to guide the instruction of individual students and to see analyses of assessment results that focus on school performance at the level of individual learning standards or performance indicators.

In their haste to meet the demands of the forces described in the previous paragraphs, testing companies, states, and the federal government are focused primarily on the attempt to make the 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, 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 and in the increased use of large-scale assessments 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, we will 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 an integrated system of assessment at the local, state, and to some extent, national levels. The proposed model is not novel, original, or innovative. In fact, with regard to 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 will be divided into three major sections. The first section will provide a definition of large-scale assessment and discuss the forces that led it to evolve to its current state. The second section will discuss the role and appropriate uses of large-scale assessments in public education. The third section will describe a comprehensive assessment system, the role that large-scale assessment plays within such a system, and how the individual components can be integrated to form a cohesive system.

Part 1: The Evolution of Large-Scale Assessment

Let's begin with the basic question: *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). As that answer indicates, the term large-scale assessment normally refers to a test rather than an assessment. Although the terms are often used interchangeably, as Wiggins (1993) explains, "the distinction between an *assessment* and a *test*…is not merely political or semantic (in that derogatory sense of hairsplitting). An assessment is a comprehensive, multi-faceted analysis of performance; it must be judgment-based and personal… An educational test, by contrast, is an "instrument," a measuring device. We construct an event to yield a measurement." The importance of the distinction will be discussed more fully in the section of the paper dealing with the role of large-scale assessment.

For the purposes of this paper, the term large-scale assessment will refer primarily to tests administered as part of statewide assessment programs. In general, those assessments consist of tests administered to all students at one or more grades at the elementary, middle, and high school levels in content areas such as English language arts, mathematics, science, and history. Those tests typically include multiple-choice items and may also include performance-based items such as short-answer items, open-ended items, and essay questions. To a lesser extent, statewide assessments may also include portfolio assessment "in which examples of students' work (essays, models, or reports) are assembled to document student progress [and performance]" (Landau et al., 1999).

A natural follow-up to our initial question (What is large-scale assessment?) is the question: *Why do we need large-scale assessment?* In an FAQ section on their company website, Measured Progress provides the following response to that question:

> In response to reform acts, accountability requirements, and the Goals 2000: Educate America Act, national agencies, states, and communities across the country have united to focus on the development of far-reaching systems that will reform educational standards and school environments. Basic reform strategies emphasize the need for high academic standards – describing what all children should know and be able to do – and high-quality assessments geared toward those standards. States are required to develop and implement assessment programs that correspond to curriculum standards and assess students in designated grade levels and subjects.

Popham (2001) provides essentially the same response with a slightly different flavor in response to the question, "what is 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. Remember, most Americans believe that the quality of education is tied directly to students' test scores. They believe that by establishing a large-scale testing program that captures a student's score on an annual basis, they have created a mechanism that will allow all interested parties to monitor the caliber of schooling delivered. (p. 34)

Popham (2001) and Landau (1999) both note that most statewide 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).

Popham dismisses discussion of the instructional contribution in the mission statement of large-scale assessment programs as nothing more than rhetoric in most cases. To some extent, he is probably correct with regard to the purpose of the large-scale assessment program. A balance scale determining 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 easily 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 made the following recommendation on standards and expectations:

> We recommend that schools, colleges, and universities adopt more rigorous and measurable standards, and higher expectations, for academic performance and student conduct, and that 4-year colleges and universities raise their requirements for admission. This will help students do their best educationally with challenging materials in an environment that supports learning and authentic accomplishment.

As one step toward implementing that recommendation, the Commission recommended

> 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.

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, standardized tests. Cannell's findings and their aftermath were summarized a decade later in Education Week's Quality Counts 1997:

> A West Virginia physician, John Jacob Cannell, attracted national attention in 1987 when he discovered the phenomenon. After learning that the majority of his state's school districts scored above average on norm-referenced standardized tests, Dr. Cannell requested test scores from all 50 states. He found that 90% of the school districts and 70% of the students across the country were "above the national average."
>
> The media jumped on the findings and on nationally normed tests, in which students are compared with a national sample of students tested in the past to establish a norm that distributes scores across a bell curve. Because testing companies tend to recalibrate the scoring only every five, six, or seven years, students tend to score better each year in the interim because their teachers get to know the questions and prepare the students for them. Eventually there emerges a situation similar to that in writer Garrison Keillor's fictional hometown, Lake Wobegon, "where all the women are strong, all the men are good looking, and all of the children are above average."

Although based primarily upon a faulty understanding of standardized testing and the interpretation of norm-referenced scores, the conclusions drawn from the Cannell study (in conjunction with the findings of subsequent studies on teachers and testing) did heighten awareness that teachers "teach to the test." In one sense, this can be interpreted as teaching directly to the content of the items included on a standardized test administered repeatedly year after year. In another sense, the phrase "teaching to the test" refers to the influence that the content and format have on teachers and students. As Wiggins (1993) explains:

> Tests *teach*. Their form and their content teach the student [and teacher] 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)

One obvious question that emerges is 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). As Wiggins (1993) describes

> Testing that teaches what we *ought* to value is technically difficult, time-consuming, and dependent upon the kinds of sophisticated task analysis that teachers have little time for. Therein lies the dilemma: few teachers and school system can imagine what a comprehensive examining system might look like – a system that gives each student many options for assessment and that "tests" the most complex aspects of performance and production. (p. 42)

The problem of a lack of statewide curriculum frameworks and standards evaporated 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 Act of 1994. (Concerns about the quality of those newly developed standards are a topic for another day.) Unfortunately, there was no corresponding seemingly "easy fix" 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 – if you can't beat them, join them. As commonly expressed, if it is a given that teachers are going to teach to the test, we will give them a test worth teaching to. Ultimately, growing

concerns about teaching to the test and a lack of confidence in teachers' ability to assess and evaluate student performance combined to result in three major changes to large-scale assessment: a heavy reliance on constructed-response items, a shift from norm-referenced to criterion-referenced test design and reporting, and a shift from an emphasis on school-level to an emphasis on student-level results and census testing.

The call for an end to a reliance on multiple-choice tests and for more "authentic" assessment of student performance resulted in several changes in the content and format of tests in the late 1980s and early 1990s such as

- Direct writing assessment in several states across the country,
- 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 CCSSO SCASS projects to develop new forms of assessments to measure high standards. (Rothman, 1995)

In Maine, the Maine Educational Assessment evolved from an assessment based largely on multiple-choice items in the mid- to late-1980s, to an assessment based on 50% multiple-choice and 50% constructed-response items in the early 1990s, to an assessment based exclusively on constructed-response items by the mid-1990s.

The shift from norm-referenced to criterion-referenced testing and/or reporting of results reflected the new emphasis on content standards (Carr and Harris, 2001). In Kentucky, the shift was complete as results on the new statewide assessment, the Kentucky Instructional Results Information System (KIRIS) were reported solely in terms of newly established performance levels. On NAEP, descriptions of student performance at various levels of proficiency based on scaled scores (e.g., 300, 350, 400) were replaced by reports of the percentage of students at or above performance standards of Basic, Proficient, and Advanced.

The emphasis on accountability for all and high standards for all combined with a lack of confidence in schools and teachers to assess students resulted in large-scale assessments becoming the primary vehicle to assess all students in the state at selected grade levels. At least once during elementary school, middle school, and high school, all students were to be measured against state standards via the statewide assessment. More than the shift to constructed-response items or criterion-referenced tests and reports, it is this change in the focus of large-scale assessments that most impacts the discussion of the future of large-scale assessment as will be discussed in parts 2 and 3 of this paper.

In addition to the obvious impact on the content of the tests, the emphasis on student-level results and the shift from tests consisting exclusively of multiple-choice items to tests containing significant numbers of constructed-response items had profound effects on large-scale assessment in terms of cost, testing time, and technical complexity. In terms of cost, it was not uncommon for statewide assessment programs to experience

500% to 1000% increases in annual cost (although statewide assessment programs still represent a very small portion of the education budget in most states). Testing time increased dramatically as students required additional time to respond to constructed-response items as well as more demanding multiple-choice items. Regarding technical complexity, there are issues that did not really exist as few as 20 years ago in areas such as scoring constructed-response items, combining scores from different item types, linking (or equating) test results across years based on relatively small pools of available items, and setting performance standards that must now be resolved in the midst of increasingly higher stakes for students and schools.

In summary, 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. Filling that void with credible results from local assessment systems will be a critical component in altering the role of large-scale assessment in public education.

**Part 2: The Role of Large-Scale Assessment**

In the previous section, two primary roles of current large-scale assessment were described – modeling and accountability. Accountability can be subdivided into the distinct categories of school accountability and student accountability. In no particular order, large-scale assessments are used:

- To serve as models of effective assessment practices for teachers and students. That is, to teach how to test (Wiggins, 1993)
- School accountability: To gauge the success of schools and school systems in order to hold educators accountable for student attainment of educational results (Landau, 1999).
- Student accountability: To measure the achievement of individual students to better inform instruction for students and/or to ensure that students possess an agreed upon set of knowledge and skills prior to promotion or graduation.

The primary question to be addressed in this question is what *should be* the role of large-scale assessment in public education. "Should", of course, is a complicated word full of twists and turns. Defining what "ought to be" is always more complex than defining what "can be" (e.g., due to considerations of cost, capacity, or technical limitations) or what "must be" (e.g., due to the requirements of NCLB). Therefore, let's begin with two slightly less complex questions. Is the current role of large-scale assessment appropriate? Is the current use of large-scale assessment the most effective and efficient use of large-scale assessment?

Previously, for the purposes of this paper, large-scale assessment was defined as statewide assessment. In one respect, therefore, the determining 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, if possible, with the role of the state. At the very least, the two should not be in conflict.

Public education is one of the few areas that most people agree that the state must play a significant role – other areas being health, security, maintaining an infrastructure, and governance. Reaching consensus on what the role of the state should be in public education, however, is not always 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 school funding lawsuit that consumed the state for a large portion of the 1990s (i.e., Claremont v. Merrill) was defining the role of the state in providing an *adequate* education to all students. In that case, the primary source document was the state constitution. Unfortunately, the state constitution offered no direct comment on large-scale assessment.

In January 2000, the Massachusetts Board of Education held an unorthodox meeting in which the members of the Board engaged in conversation about the role of the state and the role of the Board of Education in public education (Massachusetts Board of Education, 2000). At the end of the day, the Board 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.

Under creating effective schools, for example, there was general agreement that a major component in creating effective schools 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) As the following two quotes from the conversation demonstrate, however, the key is finding the balance between local autonomy and fulfilling the state's responsibility to public education:

> Now, this raises one of the underlying questions which we haven't directly addressed: What's the right balance between encouraging a restructuring of schools and encouraging development of leaders within schools while at the same time not running the schools?
>
> …There is a role to be played in doing research and communicating the results of that research, especially around effective schools. There's a lot of that out there. I'm not sure we need to reinvent this. Maybe a lot of this is just communication rather than original research. But there's that component. But then there's this other piece which is somehow creating enough flexibility within the system to all people to actually implement the kinds of changes and best practices that we are advocating or suggesting might be a good fit. This is the thing that is the most difficult nut for us to crack. It is ultimately a policy issue which may or may not be within the realm of the Board's authority to do much about other than to take a position on.

On the issue of accountability the Board expressed a similar need to define the limits of the state's role:

> On the accountability piece, I think there would at least be two subcategories. One is measuring. How do you know that you're achieving? And the other is: What do you do when schools aren't achieving? …We're going to start to identify those schools that are low performing. We are going to have review panels looking at those schools. …Some schools are going to pop up when the response from the review folks is: This school is in trouble." Which throws back to the Department and the Board the question: What are you going to do about it? … And while many of these things are very case-specific, there are some general issues that we need to start talking about in terms of parameters, the kinds of interventions that we ought to be taking, and where the Department and Board's roles begin and end.

These excerpts from the Board's conversation portray 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). Even within the maze of state and federal regulations, the day-to-day control of the school, its operations, and its success or failure is still a matter of local control and authority. Under such a system, the primary function of large-scale assessment is clear. The role of large-scale assessment is accountability – to provide the state with information on whether local districts and schools are meeting their achievement goals. It is important to note that the emphasis here is on auditing the performance of local districts and schools; not on auditing the performance of individual students. As evidenced by NCLB, however, auditing the performance of districts and schools requires auditing the performance of all significant subgroups of students within those districts and schools. Although many of those subgroups may be very small within individual districts or schools, monitoring or auditing the performance of subgroups is still a conceptually distinct task from monitoring the performance of individual students.

When the role of the assessment is defined then it is possible to direct attention to the design of the assessment – *form follows function*. An assessment designed for accountability may look quite different than an assessment designed to model effective classroom assessment and instruction. It is this type of distinction that former Assistant Secretary of Education Susan Neumann expressed when she called for states to consider a return to tests based largely on multiple-choice items (keynote address to the 2002 CCSSO Large-Scale Assessment Conference). 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 teachers and students are two distinct issues. During the 1970s and 1980s it is probably safe to conclude that it was not the intent of the state for multiple-choice tests to become the driving force in local instruction and assessment. The key question is what factors would make the dynamics between large-scale assessment and local assessment and

instruction different this time around. That is the question that will be addressed in part 3 of this paper.

As a transition to part 3, however, let's briefly address the second question posed at the beginning of this section: Is the current use of large-scale assessment the most effective and efficient use of large-scale assessment? The question can be considered with reference to the three uses of large-scale assessment described previously: school accountability, student accountability, and modeling effective assessment.

In terms of school accountability, the answer clearly is yes – school accountability is an effective and efficient use of large-scale assessment. Many of the large-scale assessments in place today were designed primarily as instruments to monitor school achievement. Additionally, school accountability is arguably one of the fundamental roles of the state in public education.

The appropriateness of the use of large-scale assessment for student accountability is not as clear-cut. As a vehicle to certify individual student's level of knowledge and skills on a well-defined body of standards (e.g., as a graduation requirement or exit exam), a large-scale assessment can be an efficient tool to separate those students who clearly possess the necessary knowledge and skills from those who clearly do not. As with any gross measuring device, however, it may not be the best alternative to make fine distinctions among students near the borderline. Consequently, such uses of large-scale assessment inevitably require multiple opportunities for measurement as well as a process to appeal decisions based on the assessment. As a tool to provide feedback and inform instruction for individual students, however, large-scale assessment is wholly inadequate. Most importantly, large-scale assessment is external and comprehensive. Consequently, it is neither embedded within instruction (i.e., administered at a proper time during the instruction-assessment loop) nor can it provide immediate, timely feedback that is an essential component of effective student assessment (Good & Grouws, 1979; Kulik and Kulik, 1979). Note that efforts to provide immediate results of large-scale assessments through the use of computers and other technology do not directly address the issue of immediate feedback to impact and improve instruction.

As a tool for modeling (and promoting) effective classroom instruction and assessment practices, large-scale assessments appear to continue to function effectively. A multi-year study conducted by the National Board on Educational Testing and Public Policy found that large-scale assessments influence teachers' instruction and assessment practices both positively and negatively. Among the positive effects cited were "… a renewed emphasis on writing, critical thinking skills, discussion, and explanation." (i.e., the types of skills that the tests were redesigned to promote). Additionally, the report indicated that the findings suggest the effects on instruction and assessment practice (both positive and negative) increase as the stakes associated with the large-scale assessment increase. (Clarke et al., 2003). As will be discussed in the following section, however, it is not clear whether the link between the large-scale test and classroom instruction/assessment is inevitable, desirable, or necessary.

Part 3: A Comprehensive Assessment System

Carr and Harris (2001) describe and advocate for a comprehensive assessment system to drive instruction at the local level. The comprehensive assessment system draws on data from the state/national level, the local 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). In such a dynamic assessment system, the role of the state/national assessment and its influence on daily classroom life would automatically be diminished. At the heart of the system, of course, is the given that quality assessment information is being generated at all levels.

When the large-scale assessment community (i.e., testing companies and states) adopted the *'let's give them a test worth teaching to'* attitude in the late 1980s and early 1990s there was a void driving teachers' predilection for mimicking standardized tests. The first component of the void was the absence of established curriculum frameworks or learning standards. The second component was the absence of any knowledge or understanding of the principles and methods of assessment. As discussed at the conclusion of part 1, the shift from multiple-choice tests to more 'authentic' or 'classroom-friendly' item types was not designed to address the causes of the void or to fill the abyss – both tasks that would require considerable time and effort beyond the scope of the assessment community. The shift was more of a temporary patch or short-term solution.

Now, it has been twenty years since the publication of *A Nation at Risk*, approximately fifteen years since the publication of the findings of Cannell and related studies, and nearly a decade since the passage of the Improving America's Schools Act of 1994. One question to address, therefore, is whether there has been any change in the underlying causes of the problem of teachers *teaching to the test*.

On the surface, 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 two main effects and an additional perceived benefit 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 the states' large-scale assessments are aligned with the curriculum standards (e.g., adequately sample the breadth and depth of the standards), the need and desire to teach to the test will be diminished. At the very least, the difference between teaching to the standards and teaching to the test should become less distinct. Even when there is no distinction, however, it may be a considerable period of time before the public, teachers, or even state departments of education make the cognitive shift from a focus on the test to a focus on the standards. Old habits are hard to break.

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-millenium symposium sponsored by the National Council on Measurement in Education (NCME), Richard Stiggins reported "the state of classroom assessment affairs is dismal. If 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." (Stiggins, 2001).

However, there are some pockets 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.

The state of Maine has designed a statewide assessment system that attempts to strike a balance between traditional large-scale assessment and strong local/classroom assessment. At the heart of the system is a Comprehensive Local Assessment System to be developed by each school unit (i.e., district) within the state. Following assessment standards incorporated into state regulations, and assessment development guidelines issued by the Department of Education, each district must develop an assessment system to measure student and school achievement of the standards contained in Maine's Learning Results at the elementary, middle, and high school levels. Corresponding to the grade spans at each of those levels (i.e., grades 4, 8, and

11) the state will continue to administer the large-scale Maine Educational Assessment to students statewide. The state will support the efforts of the local districts through professional development, management tools, and by providing model assessments that local districts can incorporate into their own systems.

Notwithstanding the gloomy history of classroom assessment, Stiggins (2001) recommended five actions to give hope to a brighter future for the use of classroom assessment.

1. Rethink our beliefs – develop assessment practices that deliver accurate information into students' hands in a timely and understandable manner.
2. Take an international perspective – (a) ensure that teachers are skilled assessors of student learning, (b) increase funding for professional development, (c) reduce obstacles, especially the influence of external tests, that dominate teachers' work.
3. Advocate balance in assessment – no single assessment can meet all needs, ensure the quality of both classroom and standardized assessment.
4. Learn about classrooms – understand the classroom assessment context
5. Team up with faculties of teacher education – typically teachers are not trained to do assessment-related activities well.

There are a number of commonalities among the model advocated by Carr and Harris, the examples provided above and Stiggins' action items. Among those are the following:

a) a balanced assessment approach in which 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 external tests will diminish.

b) the importance of immediate, accurate feedback to nourish the instruction-assessment loop. That level of feedback can only occur with local, curriculum-embedded assessments. Large-scale assessment can never provide this level or type of feedback.

c) support and training to cultivate quality assessment at the local level. As discussed previously, the use of large-scale assessments as models for effective assessment practices was at best a stopgap, not a permanent solution to improve teachers' assessment practices. The abyss in teachers' assessment-related knowledge and practices must be filled from the bottom up (i.e., through effective teacher education). It is extremely difficult to fill a hole from the top down.

d) materials to cultivate quality assessment at the local level. A critical component to the success of any efforts to enhance classroom assessment will be the availability of quality assessment materials. The state has a responsibility to provide these materials in addition to their large-scale assessment. The state of Maine recognizes this in their approach to the implementation of their Local Assessment System program.

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 assessment results, the danger is great that in the coming years we will travel so far down the large-scale assessment road that it will be 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 will be somehow inferior to large-scale assessment results. Rather, the argument is that in an ideal system there will be internal consistency among the various components.

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.

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

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.

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

Good, T. L. & Grouws, D. A. (1979). The Missouri mathematics effectiveness project. Journal of Educational Psychology, 71, 355-362.

Gullickson, A.R. (1985). Student evaluation techniques and their relationship to grade and curriculum. Journal of Educational Research, 79 (2), 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.

Kulik, J.A. & Kulik, C.C. (1979) College teaching. In P.L. Peterson & H.L. Walberg (Eds.) Research on teaching: Concepts, findings, and implications. Berkeley, California: McCutchan.

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.

Massachusetts Department of Education (2002). *Massachusetts Comprehensive Assessment System: Report of 2002 Statewide Results.* Malden, Massachusetts: Massachusetts Department of Education.

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

Measured Progress (2003). *Large-Scale Assessment: Frequently Asked Questions*. www.measuredprogress.org/ProductsandServices/Large-ScaleAssessment/FAQ.html

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

Office of Public Instruction (2001). *Release of IOWA test scores, MontCAS memo*, 8/9/2001, p. 5

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

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

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

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