Formative Reform:

Purposeful Planning for the Next Generation of Assessment and Accountability Systems

Charles A. DePascale
Center for Assessment (NCIEA)
June 2009

Fanaticism consists in redoubling your efforts when you have forgotten your aim.
(Santayana)

Introduction

It appears that we are at a crossroads in the history of public education in the United States. Factors that have been gaining momentum for several decades have reached a tipping point in 2009. Trends in early childhood education and postsecondary education are breaking the K-12 borders that have defined the limits of universal public education since the middle of the twentieth century. The global economy and interconnectedness of the *flat* world are fueling an unprecedented demand for common, national standards. Innovation in and the increased availability of technology increases the likelihood that quality instruction and highly qualified teachers will reach all communities and students across the country. For a wide variety of reasons, substantial control over school funding and policy has shifted from the local community to regional, state, and federal levels. The concern for the individual child and the education of all children that emerged in the 1960s and 1970s is now pervasive.

In the midst of this convergence of forces, the public at large and the measurement community are in the process of making policy decisions that will significantly impact public education over the next ten to twenty years. At the public level, Congress and the Administration are considering the next reauthorization of the Elementary and Secondary Education Act – making decisions regarding accountability that will directly impact the shape of public education at the state, district, and school levels. Within the measurement community, the latest revision to the Joint Standards for Educational and Psychological Testing currently under development must address an educational testing environment that had already changed significantly by the time the last version of the Standards was issued in 1999 (AERA, APA, NCME, 1999) – an environment in which measurement may not be the primary purpose of educational assessment.

As we consider the next generation of assessment and accountability systems in this *brave new world* of public education it is to our advantage to pause and engage in a process of formative education reform. That is, to define the purposes of our accountability systems and assessment systems, as well as the purposes and goals of the public education system that they are intended to support. Then within the context of those purposes and goals to process feedback from the assessment- and accountability-based reform efforts of the last 35 years to determine which aspects of the systems are

effective and which are not, the conditions and contexts under which their effectiveness is facilitated or impeded, and the adjustments to those systems that are necessary.

Therefore, our goal in this paper is to provide background information and pose questions to inform this process of purposeful reflection and decision-making. We do so through discussions of Public Education (Section 1), Assessment (Section 2), and Accountability (Section 3). Although the three areas are interrelated and interdependent, there are unique aspects and decision points within each area that should be considered and understood to build a coherent system of assessment and accountability designed to support the goals of public education.

Section 1: Public Education

- *What are the basic tenets that shape the American public education system?*
- *Which principles are sacrosanct and which are open to debate?*
- *In which ways are the principles rigid and in what ways are they sufficiently malleable to meet to the current interests, beliefs, demands, or needs of society?*
- *What aspects of American education will remain largely unchanged in the next fifteen to twenty years because they are fundamental and not because of financial, technological, political, or economic constraints?*

Any reflection on education systems in the United States must begin with the concept of public education – the cornerstone of American education. In one sense, *public* refers to the *individual* and universal access to education that is freely available, unrestricted, open, and supported by government funds. In a second sense, public refers to *society* and an education system that is for the common benefit of and is the shared responsibility of the community. The importance of finding the proper balance between the individual and society has long been part of the dialogue on public education as expressed in the Cardinal Principles of Education published in 1918.

> Consequently, education in a democracy, both within and without the school, should develop in each individual the knowledge, interests, ideals, habits, and powers whereby he will find his place [i.e., that vocation and those forms of social service in which his personality may develop and become most effective] and use that place to shape both himself and society toward ever nobler ends.

In many ways, the individual and societal facets of public education are compatible and interconnected (e.g., there is general agreement that educated citizens are necessary for a well functioning society). However, there are cases in public education (as in other publicly supported institutions) where the relationship between *individual rights* and the *greater good of society* is strained. Understanding the relationship between a community's responsibility to provide adequate access and opportunity to education and an individual's responsibility to avail oneself of the opportunities provided is also a constant struggle. Finally, it is also clear that the definition of *community* with respect to public education is evolving as the historical definition of community as local city or town has resulted in unacceptable inequities in access, opportunity, and outcomes.

Purpose

The purpose of public education is …

The manner in which one chooses to complete that sentence must have a profound impact on the design and implementation of the education system as well as on the systems and tools used to evaluate its implementation and measure its outcomes. There is little question that there are now and have been in the past differing perspectives on the purpose of public education. One broad perspective on the purpose of public education has been expressed through a description of the qualities of an educated child that has

been central to two court decisions in landmark education finance lawsuits in Kentucky and Massachusetts that shaped the current education reform movement that began in the late 1980s and early 1990s:

> [a]n educated child must possess 'at least the seven following capabilities: (i) sufficient oral and written communication skills to enable students to function in a complex and rapidly changing civilization; (ii) sufficient knowledge of economic, social, and political systems to enable students to make informed choices; (iii) sufficient understanding of governmental processes to enable the student to understand the issues that affect his or her community, state, and nation; (iv) sufficient self-knowledge and knowledge of his or her mental and physical wellness; (v) sufficient grounding in the arts to enable each student to appreciate his or her cultural and historical heritage; (vi) sufficient training or preparations for advanced training in either academic or vocational fields so as to enable each child to choose and pursue life work intelligently, and (vii) sufficient level of academic or vocational skills to enable public school students to compete favorably with their counterparts in surrounding states, in academics, or in the job market.'[1]

Within this perspective of an educated child, one can argue about which specific knowledge and skills are necessary as well as the extent to which they are necessary to reasonably claim that a child is an educated child. There is no room within this perspective, however, for the argument that an educated child can be defined or measured solely through the level of English language arts or mathematics skills attained.

Equal Access and Opportunity

Although *equality* has long been one of the fundamental principles underlying American public education, the concept has taken on new meaning in recent decades. Societal attention to civil rights for all individuals and particular groups has impacted public education greatly. Federal legislation mandated increased attention to economically disadvantaged students in the 1960s (ESEA – Chapter 1, 1965) and students with disabilities in the 1970s (P.L. 94-142, 1975). Trends in education shifted toward student-centered instruction with special emphasis in recent years on multiculturalism, social justice, and increased focus on appropriate instruction for English language learners.

This concern for equality and the education of all children is captured in terms or concepts such as "opportunity to learn" and "equal access." Arriving at a common understanding of these terms and their operational application to public education has been difficult. Two critical issues are discussed in the following sections: the distinction between adequate and equal, and the meaning of equal outcomes.

[1] McDuffy v. Secretary of the Executive Office of Education, 415 Mass. At 618-19, 615 N.E.2d (quoting Rose v. Council for Better Education Inc.. S.W.2d186, 212 (Ky, 1989)) downloaded from the Massachusetts Department of Education website: Education Laws and Regulations, The State Constitutional Mandate for Education. www.doe.mass.edu/lawsregs/litigation/mcduffy_hancock.html.

Adequate v. Equal

The issue of equal access and opportunity to public education has been played out in the courts and has been framed to a large extent in terms of funding and related financial issues. In some states, truly local funding of public education has been replaced by systems in which a significant portion of local revenues for education are collected by the state and redistributed to local cities and towns in processes designed to equalize funding[2] across school districts. Local independence regarding the use of funds that are received for education has also been diminished by numerous state and federal regulations related to access and opportunity for protected groups of students as well as spending requirements attached to federal and state funds distributed to local school districts.

In New Hampshire, a series of state supreme court cases on school funding in the 1990s (Claremont School Dist. v. Governor, 1993, 1997, 1999, 2002) began with the basic issue of equal funding across school districts, shifted in focus to the role and responsibility of local communities to pay for and/or provide special services to students with disabilities, and ultimately evolved into an ongoing decade-long debate over the question,

> Does a commitment to public education ensure each child access to *equal* educational opportunities or simply to *adequate* educational opportunities?

The question arose in New Hampshire due to wrangling over interpretation of wording in the state constitution, but it raises fundamental issues related to equal access and opportunity for all students, local control of education, and the purpose of a public education. Can educational opportunities be adequate and unequal? Is it possible to provide students with access to educational opportunities that are sufficient or good enough to meet some established marker (e.g., proficiency on the state standards) if those opportunities are not equal to those received by other students within the same school, district, state, or across the country?

The already complex question becomes more complicated and controversial when framed in this way:

> *Is it acceptable for a community to determine that it wants to offer more than the state established minimum or norm in its public schools?*

Under the conception of equal access and opportunity to public education is a local community allowed to raise supplemental funds for education for the purpose of improving education for their students through efforts such as

a) attracting better teachers through higher salaries and/or benefits,
b) improving facilities,
c) increasing resources such as technology and books,
d) increasing the range of extracurricular activities available to students, or
e) increasing additional supports to students through after school, weekend, or summer programs?

If such efforts to provide improved opportunities are not appropriate within the context of the public schools is it acceptable for the community to offer those opportunities within

[2] Note that equalizing funding may or may not mean that each community receives the same amount of funding.

other contexts such as community service departments, public libraries, or youth groups? To push the question from simply unanswerable to ridiculous is it acceptable for private organizations within a community, groups of parents, or even individual parents to offer particular students access to additional or better educational opportunities?

Equal Outcomes

At some point, the issue of equal access and opportunity shifts from inputs to outcomes. It has been clearly established that equal access and opportunity for all students does not mean that all students are provided access to the *same* educational opportunities. Rather, equal access and opportunity means that each student is provided access to *appropriate* educational opportunities for her or him to achieve certain goals or outcomes. When the focus shifts to outcomes, however, another set of questions emerges.

Is it the purpose or goal of public education to ensure that all students attain the same educational outcomes? If so, what does that mean?

a) Do we expect that there will be no variation in achievement among individual students?
b) Do we expect that there will be no mean difference in achievement between particular subgroups of students or no correlation between educational achievement and group membership or status?
c) Do we expect that there may be variation among individuals or mean differences among groups, but that all students will reach an established and accepted minimum level of achievement?

If we do expect all students to attain the same minimum level of achievement within a particular content area

a) How do we determine the appropriate level of achievement?
b) Do we expect all students to reach that level at the same time or with the same amount of effort?
c) Do we expect all students to follow the same pathway to reach that level?
d) Do we expect all students to be able to demonstrate their level of achievement in the same manner?
e) Do we expect students to be able to demonstrate their level of achievement in multiple ways within the same context? In one way across multiple contexts? Across multiple ways across multiple constructs?

The education systems that we design should vary tremendously based on our answers to those questions.

Structures

The current education system is largely defined by structures. These structures include the physical structures or school buildings which through their design and location both reflect and dictate educational policy. Perhaps at least as significantly, the education system is defined by organizational and political structures that dictate matters such as

a) the relationship between schools and the greater community,
b) the governance of schools by external, community-based school boards,

c) the organizational structure of districts and schools both in terms of functional units and collective bargaining units,
d) the organization of the school day and school year,
e) the education, training, and selection of teachers, and
f) the beginning and end points of public education.

Through much of the second half of the twentieth century much of the focus of public education has been on K-12 education. Even as the G.I. Bill[3] opened college education to a wider pool of students after World War II and there was a dramatic increase the numbers and percentages of students attending some type of postsecondary institution there has been a clear demarcation between K-12 and higher education, and a prevailing view of high school as the end point of the community's responsibility to provide a public education. Correspondingly, across most of the country kindergarten remained the traditional latest starting point to public education even as the importance of pre-school and early childhood education became widely acknowledged and accepted. As we enter the early years of the twenty-first century, the push for public education to break through both the kindergarten and grade 12 barriers seems to be gaining momentum. Early childhood education is a top priority at the federal level and in many states. The President has declared postsecondary education a matter of national security and called for all Americans to commit to at least one year of some form of postsecondary education or training.

As we grapple with defining the purpose and scope of public education at the beginning of the twenty-first century, however, it will do us well to remember that such debate has been a hallmark of American public education. Since the inception of public high schools in the early 1800s to their national growth following the Civil War there has been ongoing questioning about their purpose, whom they should serve, what subjects should be offered, their relationship with colleges, and the preparation and development of their teachers. Much of the current debate about public education can be well informed by a review of three key historical documents that helped shape American public education:

- The Comprehensive High School – a series of reports by James Conant between 1959 and 1967 on a study sponsored by the National Association of Secondary School Principals.
- The Cardinal Principles of Education – a report issued in 1918 by the Commission on the Reorganization of Secondary Education, a commission appointed by the National Education Association in 1913 to determine "those fundamental principles that will be most helpful in directing secondary education."
- The Committee of Ten Report issued in 1893 to the National Council of Education. The Committee, appointed by the National Education Association in 1892 was charged with convening committees of secondary school and college teachers in a variety of content areas to examine issues such as standards, best practices in instruction, time allotments, and best methods for testing within each content area.

[3] Servicemen's Readjustment Act of 1944, P.L. 78-346

The Comprehensive High School (1959)

In this study of American secondary education, Conant defines the comprehensive high school that was taking shape in the period following World War II and to a great extent exists, in name at least, to this day. Perhaps somewhat surprisingly, in his recommendations and findings he was not describing the one size fits all, impersonal comprehensive high schools that are decried today. The first two of Conant's 21 recommendations for the American high school call for 1) counseling systems that begin in the elementary school and include good articulation between elementary, junior, and senior high schools, and 2) individualized programs of study for every student. Additional recommendations address the need for literacy programs (including improved reading instruction at grades K-8), required and elective programs, and provide cautions about the misuse of tests. In the foreword to Conant's first report on the comprehensive high school, John Gardner describes the comprehensive high school as "a peculiarly American phenomenon."

> It is called comprehensive because it offers, under one administration and under one roof (or series of roofs), secondary education for almost all the high school age children of one town or neighborhood. It is responsible for educating the boy who will be an atomic scientist and the girl who will marry at eighteen; the prospective captain of a ship and the future captain of industry. It is responsible for educating the bright and the not so bright children with different vocational and professional ambitions and with various motivations. It is responsible, in sum, for providing good and appropriate education, both academic and vocational, for all young people within a democratic environment which the American people believe serves the principles they cherish. (1959, pp. ix-x)

The Cardinal Principles of Education (1918)

In describing seven cardinal principles, the Commission considered key objectives that "should guide education in a democracy." The principles go well beyond defining knowledge and skills in specific content areas and address academic skills as "tools in the affairs of life… [and] command of these fundamental processes, while not an end in itself, is nevertheless an indispensable objective." Several of the skills identified as key objectives of education in 1918 bear a striking resemblance to the twenty-first century skills being emphasized today and to the qualities of an educated child previously cited. The Commission listed the principles in seven named categories, but made it clear that these were cross-cutting principles that must be reflected across the academic and nonacademic programs within the school. The seven principles identified by the Commission as the main objectives of education were

1. Health – provide health instruction, inculcate health habits, organize an effective program of physical activities, regard health needs in planning work and play, and cooperate with home and community in safeguarding and promoting health interests.
2. Command of fundamental processes (reading, writing, arithmetical computations, and the elements of oral and written expression) – Throughout the secondary school, instruction and practice must go hand and hand, but …only so much theory should be taught at any one time as will show results in practice.

3. Worthy home membership – the development of those qualities that make the individual a worthy member of a family, both contributing to and deriving benefit from that membership.
4. Vocation – to equip the individual to secure a livelihood for himself and those dependent upon him, to serve society well through his vocation, to maintain the right relationships toward his fellow workers ad society, and as far as possible, to find in that vocation his own best development. This ideal demands that the pupil explore his own capacities and aptitudes, and make a survey of the world's work, to the end that he may select his vocation wisely. Hence, an effective program of vocational guidance in the secondary schools is essential.
5. Citizenship – to develop in the individual those qualities whereby he will act well his part as a member of neighborhood, town or city, state, and nation, and give him a basis for understanding international problems.
6. Worthy use of leisure – equip the individual to secure from his leisure the recreation of body, mind, and spirit, and the enrichment of his personality. This objective calls for the ability to utilize the common means of enjoyment, such as music, art, literature, drama, and social intercourse, together with the fostering in each individual one or more special vocational interests.
7. Ethical character – In a democratic society ethical character becomes paramount among the objectives of the secondary school; the development of on the part of pupils of the sense of personal responsibility and initiative and, above all, the spirit of service and the principles of true democracy which should permeate the entire school – principal, teachers, and pupils.

The Committee of Ten (1893)

The guiding questions considered by each of the content committees convened by the Committee of Ten in 1893 are listed below. With slight adjustments in terminology, the same questions could be used today to guide the current discussions regarding common standards, college readiness, and comprehensive assessment systems.

1. In the school course of study extending approximately from the age of six years to eighteen years—a course including the periods of both elementary and secondary instruction—at what age should the study which is the subject of the Conference [content area] be first introduced?
2. After it is introduced, how many hours a week for how many years should be devoted to it?
3. How many hours a week for how many years should be devoted to it during the last four years of the complete course; that is, during the ordinary high school period?
4. What topics, or parts, of the subject may reasonably be covered during the whole course?
5. What topics, or parts, of the subject may best be reserved for the last four years?
6. In what form and to what extent should the subject enter into college requirements for admission? Such questions as the sufficiency of translation at sight as a test of knowledge of a language, or the superiority of a laboratory examination in a scientific

subject to a written examination on a text-book, are intended to be suggested under this head by the phrase "in what form."
7. Should the subject be treated differently for pupils who are going to college, for those who are going to a scientific school, and for those who, presumably, are going to neither?
8. At what stage should this differentiation begin, if any be recommended?
9. Can any description be given of the best method of teaching this subject throughout the school course?
10. Can any description be given of the best mode of testing attainments in this subject at college admission examinations?

In addition to specific recommendations regarding each content area, the report of the Committee of Ten also contained a series of general recommendations that are also relevant to the current conversation:

a) Make the early years of secondary instruction as representative as possible, postponing *bifurcation* – choices between specialized programs of study – until the student has had the opportunity to experience the content areas and exhibit their skills within each area.
b) Establish clear articulation between secondary schools and colleges so that any students successfully completing an established secondary course of study would have access to postsecondary education. Completing a secondary course of study would establish that the student has spent four year studying a few subjects thoroughly; "and, on the theory that all the subjects are to be considered equivalent in educational rank for the purposes of admission to college, it would make no difference which subjects he had chosen… - he would have had four years of strong and effective mental training."
c) "More highly trained teachers will be needed than are now ordinarily to be found for the service of the elementary and secondary schools." One option to accomplish this is to provide tuition and expenses for teachers willing to devote half of their summer vacation to study at local universities. A second option is for colleges and universities to more closely align their courses to the needs of their local public schools and teachers.
d) In every sufficiently large school system, the best teacher in each department should be "enabled to give part of his time to helping the other teachers by inspecting and criticizing their work, and showing them, both by precept and example, how to do it better."

Section 2: Assessment

For most intents and purposes, in public education the term assessment has come to mean standardized, large-scale testing. There is no question that the impact of large-scale testing has increased dramatically since the passage of the No Child Left Behind Act of 2001. However, large-scale testing, predominantly in the form of standardized, multiple-choice, norm-referenced tests has long been a staple of public education at the local school, district, and/or state levels. In terms of its scope and the stakes associated with its use, the influence of large-scale testing on public education in the United States appears to be stronger now than at any time in history. The *rise* of large-scale standardized testing in K-12 public education can be traced through its use as a selection and evaluation tool for Title 1 (née Chapter 1) programs under successive reauthorizations of the Elementary and Secondary Education Act of 1965, through the minimum competency and basic skills testing programs of the 1970s and 1980s, to the growth of custom state assessments in the 1980s and the standards-based reform movement that led to innovative state assessment programs in states such as Vermont, Kentucky, and Maryland (Rothman, 1995), to the state standards and assessment requirements of the Improving America's Schools Act of 1994, and, ultimately, to the increased district/school accountability and technical oversight requirements of No Child Left Behind. With regard to high-stakes for students, the large-scale testing movement appears to have reached a high-water mark (for the time being) with approximately half of the states including performance on a standardized test among their requirements for high school graduation[4].

Concurrent with the increased reliance on standardized testing since the 1970s have been developments which have placed additional demands on large-scale testing. In this period there has been a continuation of an evolving shift from a system- or teacher-centered view of education to a more individual, student-centered perspective; an increased awareness of equity for traditionally disadvantaged groups of students; and increased requirements for educating all students – particularly students with disabilities. These factors have resulted in an expectation that large-scale testing programs will include virtually all students and will provide *diagnostic* results for all students. In many ways it may appear that the last forty years have been a golden age for large-scale testing.

In retrospect, however, it appears that the zenith of the golden age of large-scale testing actually occurred in mid-1980s. The increased emphasis on high-stakes accountability and individual student performance effectively brought an end to a measurement-driven era of large-scale testing in K-12 public education. In terms of instruments, this era was characterized by the development and implementation of large-scale assessment tools such as the National Assessment of Educational Progress (NAEP), NAEP-like assessments at the state level, and the nascent field of adaptive testing utilizing newly developing computer technology. In terms of theory, this period was characterized by the an interest in defining criterion-referenced measurement (in contrast and in relation to

[4] The measurement of achievement on college readiness standards for all exiting high school students could be the next major milestone in the system.

norm-referenced measurement) and the emergence of item response theory and its *application to practical testing problems*[5] (Bock, Mislevy, and Woodson, 1982).

This is not to argue that there have been no advances in large-scale testing since the mid-1980s. Without question, there have been significant advances in areas such as the scoring of constructed-response items, item development, alignment of tests to content standards, and setting of performance or achievement standards. The technological revolution that has occurred since the 1980s has led to advances in the use of technology, hardware and software, in many areas of large-scale testing including item development, administration, scoring, and reporting. There have undoubtedly been advances also in the use of large-scale testing to provide more accurate estimates of individual student performance.

Since the mid-1980s, the emphasis on the development of large-scale testing has increasingly shifted to its use as an accountability tool (i.e., a hammer) or lever for political and social change rather than as a measurement tool. The impact of the shift from viewing assessment, in general, and large-scale assessment, in particular, as a measurement instrument to a policy instrument and lever of political change is profound and cannot be underestimated (Baker, 1988). The primary use of large-scale assessment shifted in the 1970s and 1980s from determining *how much mathematics Johnny knows* (whether compared to a content criterion or other students) to attempting to categorize the quality of Johnny's school, teachers, or curriculum. There was also an increased emphasis on the large-scale assessment as a driver of reform and system change through its use as an accountability tool, and also as a vehicle for professional development (Baker, 1988). The terms assessment-driven reform and assessment-driven instruction that became prevalent are indicative of this shift in emphasis. This shift negatively impacted the traditional use of large-scale assessments as an indirect measurement instrument, and also provided an overly optimistic projection of the depth of instructional change that could occur in the name of education reform.

The period from the late 1980s through the 1990s also saw a shift in traditional conceptions of validity (Messick, 1989), the impact of which on the use and interpretation of large-scale testing is still slowly evolving. Tthe shift in purpose of large-scale assessments should have resulted in a corresponding shift in the validity framework used to evaluate those assessments. However, the overlap of a) the change in the purpose of assessments, b) the backlash against norm-referenced, multiple-choice tests and c) the re-conceptualizing of validity may have confounded the discussion of validity. Similarly, a second shift in the 1990s that viewed large-scale assessments as tools to simultaneously certify individual student performance (and later diagnose individual student performance) and also serve as policy indicator of school quality should have resulted in another change in the validity framework.

The overreliance on assessments as a political tool has rendered us slow to understand what large-scale testing can and cannot do. It has led us to devote too many resources to answering questions such as "How can we include more students in large-scale

[5] Reference to Lord (1980). Applications of Item Response Theory to Practical Testing Problems.

assessment systems?" and "How can we get more information about individual students out of large-scale assessment systems?" rather than asking questions such as:

- What is the best approach or best tool to gather information about what individual students know and can do – particularly students with disabilities and English-language learners?
- What is the most efficient and effective use of large-scale testing in the K-12 setting?

Only when we are able to acknowledge that large-scale assessment is not the answer to all of our needs will we be able to a) find the appropriate tools and systems to provide those answers and b) focus on making real improvements to large-scale assessments.

More importantly, the overreliance on large-scale testing for accountability has drawn attention from and slowed advances in other forms of assessment and other levels of assessment. Areas such as classroom assessment (for formative and summative purposes), performance-based assessment, computer-adaptive testing are either examined through the same lens as high-stakes testing or considered in terms of the information that they can provide to support high-stakes testing. Until recently, scant attention has been paid to the type of information that these assessments can provide, their appropriate use, and the critical ways that they differ from large-scale testing.

Matching the purpose and the method

In most endeavors matching purpose and method (i.e., finding the right tool for the job) is the most effective and most efficient approach to accomplishing a goal. In educational testing, with increasing frequency our tool box is being referred to by terms such as a "comprehensive assessment system" or "balanced assessment system." The tools within a comprehensive assessment system are referred to as traditional large-scale assessment, interim or benchmark assessment, and formative assessment.

At this point, our understanding of the purpose and appropriate use of each of the tools is nebulous, at best. After several attempts and several years the field is reaching some agreement on the definition of each term and making necessary distinctions between the instrument itself and the use of that instrument. One can only hope that investing so much time in defining the terms will make it easier to accomplish such tasks as a) identifying the appropriate use of each type of assessment, b) developing specifications for the development, implementation, and use of each type of assessment approach, and c) developing standards and procedures for evaluating each type of assessment approach.

Moving beyond assessment as a test

In education, we are so firmly entrenched within the standardized, large-scale testing paradigm that we appear to have lost sight of the important distinctions in measurement between the process of measuring, measurement tools, and the trait being measured. The term *assessment* has come to be synonymous with a test and, all too often, the trait being measured and/or the attainment of that trait is defined solely by the result of that test. Even within the context of balanced or comprehensive assessment systems, the focus of discussion is on a series of test administrations (e.g., summative, benchmark/interim, formative). Placing primary and inordinate attention on the measurement tool (i.e., the

test) increases the danger of limiting consideration of the purpose and process of measurement and the traits being measured to those that can be addressed with that tool. Rather than first identifying essential knowledge and skills or content and processes and then determining the most appropriate measurement approaches and instruments, we run the risk of measuring only the content, knowledge, skills, etc. that can be easily measured on a standardized, large-scale test.

Additionally, when all assessment instruments are examined within the framework of large-scale testing, they are expected to possess the same technical characteristics as large-scale tests. Although technical qualities such as reliability and validity are desirable, when they are defined by statistics developed for large-scale assessments there are two major areas of concern. First, reliance on traditional statistics (e.g., Cronbach's coefficient alpha to measure reliability) can narrow our conceptions of a technical quality such as reliability to those that were developed for and can be appropriately applied to standardized, large-scale tests (Moss, 1994). Second, reification and deification of such statistics as indicators of technical quality can divert attention from the purpose of the assessment and the actual content and skills being assessed.

To make full use of assessment we need to move beyond the constraints imposed by a view of assessment centered on standardized, large-scale testing.

1. What are the proper methods for evaluating classroom assessment instruments of various types including tools such as observations and checklists within the contexts of formative assessment, classroom grading, and/or school accountability?
2. What statistics (parametric and non-parametric) are appropriate to examine the technical characteristics and quality of assessment instruments designed to focus on individual student performance rather than group performance or performance across several points in time rather than a single point in time?
3. In what ways can our concept of proficiency be improved and made more useful by the understanding that it must exist and be measurable outside of a score on a large-scale assessment instrument?
4. How will a better understanding of the distinctions between assessment as an evaluation process, assessment instruments as measurement tools, and the outcomes of the assessment impact the design and use of assessment in education?

Moving beyond the test as an accountability tool

We frequently speak of using assessment at the state and federal level to improve instruction or inform instruction. Stiggins and others have persistently attempted to make the distinction between assessment *of* learning and assessment *for* learning at the classroom level. In spite of the common use of terms such as these, however, we continue to view tests as an accountability tool and design tests to provide information for accountability purposes. That is, we continue to design tests which yield a single, reliable test score that is an aggregate, or composite, score reflecting student performance across a domain.

In the case of the large-scale testing movement that emerged in the late 1980s and early 1990s there was a direct attempt to improve instruction through accountability and large-scale assessment. Based on the perception that classroom instruction and assessment were heavily influenced by large-scale assessment for accountability there was a conscious attempt to improve instruction by developing better large-scale assessments and holding schools accountable for results on those assessments. (Linn, 1994). Adopting an "if you can't beat them, join them" attitude the assessment community and policy makers decided that if classroom instruction and assessment were going to be modeled after large-scale assessment then we were going to "make assessments worth teaching to." (Resnick & Resnick, 1992, p. 59) As one component of a multi-pronged effort to improve instruction, the decisions to better align large-scale assessments for accountability with classroom instruction and to model alternative and appropriate assessment techniques through large-scale assessment have merit. As the primary or sole component of the effort to improve instruction those actions were much too naïve, one-dimensional, and did nothing to address the corrupting impact that accountability can have on instruction and assessment. (Baker, 1988; Linn, 1994). Additionally, those approaches do not address the fundamental question of whether measuring what students know and can do – the primary purpose of assessments for accountability – is sufficient information to inform and improve instruction.

There is no question that for accountability purposes most assessment instruments are designed to provide information on what students know[6]. There is a question, however, about whether information on what students know is sufficient to inform and improve instruction. Pellegrino et al. (2001), in *Knowing What Students Know*, draw a connection between assessment and instruction through their *assessment triangle* of cognition, observation, and interpretation where *understanding* what students know is critical to informing and improving instruction. That level of understanding includes not only knowing what students know, but also a) knowing how and why they arrived at their current level of knowledge and skills, and b) knowing what they do not know and why. Using that knowledge to improve instruction requires a deep understanding of a) the knowledge and skills that students must have to be successful, and b) knowing how to move students from their current level of knowledge and skills to the necessary and required level. Realizing that there is a distinction between the type and level of information provided by assessment instruments designed for accountability purposes and the information needed to improve and information is crucial to the development of appropriate assessments (Mislevy, 1996), and is a critical first step in moving beyond the test as an accountability tool.

Fortunately, there is a long history in educational measurement of attempting to build assessments to provide better information about what individual students know and do not know. The Holy Grail of testing is the test or test score that indicates with certainty that a student possesses all of the knowledge and skills up to a certain level within the domain being measured and nothing beyond that level. On a more realistic level, there

[6] Issues related to how well current assessments accomplish even that purpose, their level of accuracy, and degree of precision are worthy of debate, but are secondary to the argument in this section.

are examples of successful efforts to identify learning progressions within narrowly-defined domains (Forster and Masters, 2004; Wilson and Draney, 2004) and efforts to identify students' misunderstanding and misconceptions through carefully developed assessment items (Petit, 2009). Although these efforts may have been operating in the shadows of large-scale, standardized assessment they are ongoing and there is a solid foundation to build on in this area.

Unfortunately, the development of tests that provide information to support a better understanding of what students know is, to a large extent, outside of the realm of current test theory and item response theory techniques. The theory and techniques supporting it are based largely on an overall and unidimensional conception of proficiency, and are designed to describe typical, group-level performance. As Mislevy (1996) explains,

> The IRT model does not address the reasons that some items might be more or less difficult than others.… Now from a cognitive perspective, what makes a task difficult for a particular individual is the matchup between her knowledge structure and the demands of the task. As noted above, these matchups vary from one person to another for any given task. An IRT item difficulty parameter captures only the relative ordering of items *on the average*. The summaries of the difficulties of items and the proficiencies of persons that the IRT parameters embody will therefore forego potential information in any given person's responses to the extent that items are hard for some people and easy for others. (pp. 393-394).

Or stated more simply,

> A coarser grain size may well suffice for accountability purposes…The grain size of the … proficiency guidelines, for example, serves well for summary indicators of learning to monitor progress, but is too coarse for specific instructional guidance. Two "mid-novice" students might require quite different experiences to progress to "high novice." (p. 391).

Attempts to develop and implement multi-dimensional IRT models, if successful, might provide finer information about what students know, but may still not address the question of why and how a particular student knows what she knows and what instruction the student needs to progress. And, as suggested in the previous section of this document, breaking free of the constraints of current practices and techniques may be a difficult task.

Section 3: Accountability

In the last decade we have expended a lot of time and energy trying to determine *how* to develop accountability systems based on large-scale assessments results to meet the criteria established under No Child Left Behind. While becoming tangled in the mechanics of those accountability systems we have accepted without significant public debate the more critical questions of *who* we are holding accountable, for *what* we are holding them accountable, and *why* we are holding them accountable. There has been criticism and commentary on these issues, to be sure, but not the level of open public debate that leads to

a) refinement and clarification of goals and purposes, then to
b) deeper understanding of the goals, and ultimately to
c) acceptance, buy-in, and commitment to meet the goals.

Only when those issues are resolved can we better focus on questions related to how to develop accountability systems that best support our goals.

1. What other options are available beyond test-based accountability?

2. What key outcomes or processes do we want to include in accountability systems that cannot be measured through large-scale assessments of student achievement in mathematics and English language arts?

Why accountability?

It is reasonable to assume that there is a rationale and purpose behind the development and implementation of an accountability system. That is, one expects the act of implementing the accountability system to accomplish something tangible. In the case of federally funded Title 1 programs under NCLB, two possible tangible outcomes of the accountability system are listed below.

- Outcome A: States' programs are not funded if they do not meet their contracted purpose to provide instruction and supplemental services needed so that underperforming students who are economically disadvantaged meet grade level academic achievement standards in reading and mathematics.
- Outcome B: Underperforming students who are economically disadvantaged meet grade level academic achievement standards in reading and mathematics.

Outcome B is the more expansive of the two outcomes listed and it directly links the accountability system and the broadly defined purpose of Title 1 – to improve the education of economically disadvantaged students. Outcome A, on the other hand, is more closely aligned to the operational goal of the federal Title 1 effort – to fund state programs that improve the education of economically disadvantaged students.
Consider the following questions:

a) Who is being held accountable and for what are they being held accountable under Outcome A and Outcome B?
b) Is one of the two outcomes listed more appropriate within our context?
c) If both outcomes are important can they be served by a single accountability system or would an accountability system designed to meet Outcome A be significantly different that a system designed to meet Outcome B?

d) Are current accountability system better designed to meet Outcome A or Outcome B?

The rhetoric associated with accountability systems clearly indicates that the prevailing view is that implementing an accountability system will result in the ultimate goal (e.g., improved learning). The purposes of the accountability system are indistinguishable from the goals of a broader program that the accountability system may be intended to support Braun, 2008). In fact, it seems increasingly likely that there is no program behind the accountability system, but merely a policy[7]. In many ways, the accountability-driven reform philosophy of NCLB is merely the *logical* next step following the assessment-driven reform movement of the 1980s and assessment-driven instruction and reform movements of the 1990s.

Who is accountable?

Explicit in the term *accountable* is the expectation that some person, persons, or organization is going to be required to explain their actions or results, at a minimum, or perhaps face consequences based on their actions or results. Implicit in the term is the expectation that the persons being held accountable have the ability to impact those actions and results for which they are being held accountable. Consequently the effectiveness of an accountability system will be limited by the extent to which it a) does not actually hold someone accountable or b) holds people or organizations accountable for things over which they have little or no control.

Accountability systems developed for NCLB are designed to hold a school, district, or state accountable for student results on statewide mathematics and reading tests. In short, there is a relatively narrow focus with a vague sense of responsibility. At the school and district level, results from NCLB accountability systems do not provide sufficiently precise information to identify the specific person or persons responsible for the results nor do they provide information that can be used to apportion responsibility for positive or negative results among the district superintendent, curriculum coordinator, principal, teachers, education specialists, or any other staff members. What is the impact on the effectiveness of an accountability system when everyone is accountable for the overall results, but no one person is accountable for any specific portion of the results?

One consequence of the vague sense of responsibility associated with an accountability system commonly labeled a *district* or *school* system is that it tends to focus attention and actions on the narrowly defined school organization (e.g., principal, teachers, and other school staff) rather than the broader community responsible for public education. An accountability system that allows responsibility to be focused narrowly within the school building does nothing to increase the sense of community responsibility for public education.

[7] We are making a distinction here between program elements directly related to the accountability system (e.g., technical characteristics of the assessment, participation requirements, reporting requirements) and program elements directly related to improving instruction and learning at the school level. Some may argue that the distinction is blurred and that accountability system requirements such as participation by all students (including those with significant disabilities) in the accountability and assessment systems is a program decision that directly impacts instruction and student learning.

By design, in our system of public education the community at large, the local school board, parents, and students all have an active role in the success of the education system. Several external factors, however, are weakening the sense of community ownership and responsibility for public education:

a) an aging and mobile population in which a dwindling portion of the community has a direct connection with the public school system,
b) lessening of direct financial control and responsibility when the funding community has been more broadly defined to the state and federal level, and
c) lessening of direct policy control due to increased federal and state regulations governing all aspects of the system from the school building through instruction of individual students.

The extent that these factors and others such as the regionalization of school districts or statewide teacher contracts change the role of the community in public education must be understood in the design of accountability systems.

For what are they accountable?

Current school accountability systems focus almost exclusively on student performance in reading and mathematics. None of the other academic content areas routinely included in the elementary or secondary curriculum are included in the accountability system. None of the other major goals and purposes of public education are included in the accountability system. What is the impact on public education, or any system, when an accountability system is designed to hold people accountable for only a small portion of their responsibilities? The logical response is that all of the system's energy and resources will be shifted toward meeting the goals for which they are being held accountable and away from the other goals and purposes (Rothstein, 2009). That is not to argue, however, that the current focus on reading and mathematics under NCLB must be inappropriate.

Our current school accountability systems are housed within Title 1 of NCLB – the most recent reauthorization of the Elementary and Secondary Education Act of 1965. For more than 40 years, a primary focus of Title 1 has been on improving reading and mathematics performance for particular subgroups of students. It is logical that an accountability system within the context of Title 1 would focus on those goals. It is also fair, however, to question whether the explicitly or implicitly stated goals of Title 1 have broadened during the last 40 years and/or whether the design of the current accountability system is aligned with the goals and purposes of Title 1. It is also fair to question the consequences of an accountability system aligned to the narrow goals of Title 1 in the absence of other accountability systems of equal stature aligned to other goals.

In addition to alignment with the goals of Title 1, there are other reasons why it may be appropriate to focus an accountability system narrowly on reading and mathematics performance. Each of these reasons listed below (and others) can be fully discussed and either dismissed or accepted in an open debate on the design and purpose of an accountability system.

a) Reading and mathematics performance (i.e., literacy and numeracy) are the most important goals of public education.
b) None of the other goals or purposes of public education can be met unless or until there is adequate reading and mathematics performance.
c) If the goals in reading and mathematics performance are met, it is likely that all other goals will be met as well.
d) Reading and mathematics performance are the goals over which districts and schools have the most direct control.
e) All of the other important goals of public education are being met and do not require an external accountability system.
f) There are other internal and/or external accountability systems in place to monitor the other goals of public education.
g) All of the other important goals of public education are more difficult to monitor and measure than reading and mathematics performance, and therefore, cannot be included in an accountability system.

Other outcomes, inputs, and processes

In the opening paragraphs of this section we discussed the question *for what are they accountable* in the broad context of which of the major goals of public education can and should be included in an accountability system. Within or across goal areas the question can also be asked in terms of the level of outcomes, inputs, or processes for which districts and schools are held accountable – that is, what is actually measured by the accountability system. To what extent does the choice of what is measured by the accountability system impact its results, the validity of those results, and ultimately, its contribution toward the goals of the program?

Although it may seem counterintuitive given the reams of paper devoted to regulations, guidance, multiple versions of states' accountability workbooks, and states' submission of evidence to peer review, accountability systems developed for NCLB are stark. Virtually all decisions about district and school *quality* are based on two proficient/not proficient decisions determined from the results of a single reading test and a single mathematics test administered at each grade level[8].

No other indicators of student outcomes in reading and mathematics are included in the system. There is no attempt to collect multiple measures of student performance within a single context, to measure student performance across multiple contexts, or even to determine typical student performance as opposed to a single snapshot of performance at a fixed point in time. In the broad context of improved student learning, are we collecting sufficient information about student performance in mathematics and reading to make the required proficiency decisions about individual students or groups of students within a school?

[8] Other outcomes such as attendance rate and high school graduation rate are nominally included in the accountability system, but can be controlled so that they have little impact on results. Participation rate is a compliance indicator that is an important component of the system, but not directly related to proficiency.

Beyond outcome indicators, no other information on inputs or processes that are likely to impact performance in reading and mathematics are included in the system. In Testing, Teaching, and Learning (NRC, 1999), a pre-NCLB report issued by the National Research Council's Committee on Title I Testing and Assessment, acknowledges "although the theory of standards-based reform placed great emphasis on what students should know and be able to do, it remained silent about the knowledge and skills needed for teachers." (p. 74). The committee concluded that districts and schools lack the capacity to monitor and analyze instruction on their own. Therefore, "examining instructional practices, along with data on performance, and using that information to develop a professional development strategy, can help teachers improve their instruction and help improve student performance." (p. 76).

At the 2008 National Conference on Student Assessment, Jim Ysseldyke presented a list of 10 inputs and processes that he described as "major instructional factors" based on research on effective teaching. The final three factors (indicated by an *) were described as "especially critical factors" related to effective teaching (Ysseldyke, 2008).

1. Instructional match
2. Feedback
3. Monitoring of student performance and understanding
4. Cooperative learning
5. Relevant practice (guided then independent)
6. Personalized instruction
7. Adaptive Instruction
8. Differentiated Instruction*
9. Academic engaged time*
10. Progress Monitoring (IF used to adapt instruction)*

Is there any place in an accountability system for measures of school performance on factors such as these? Would they provide information that would directly impact our judgment of school quality? Would they provide information that would help to explain school performance and be useful to schools in need of improvement? With so much research devoted to effective teaching and instruction are there clear cut factors, and performance thresholds within those factors, that are solid indicators of school performance?

The discussion in this section focused narrowly on the issue of additional outcomes, inputs, and processes. Only additional measures of mathematics and reading performance and research-based indicators of effective instruction were considered. We avoided the controversial discussion of the inclusion in an accountability system of factors that might be used to account for low levels of performance and mitigate consequences associated with that performance. In a full and open discussion of accountability systems and school quality, however, the implications of including or excluding such factors would have to be considered.

Validation

Finally, any reflection on the purpose of accountability systems would not be complete without consideration of validity and a validation plan. Under NCLB, considerable attention is given (in theory or in regulations) to the validity of the assessments included in the accountability system. Much less attention is devoted to the validity of the accountability systems based on the results of those assessments. Having defined the purpose of the accountability system and developed an accountability system designed to accomplish that purpose, how do we judge the validity of the accountability system? How do we know that the "right schools are being identified" by the system and that student learning is improving as a result of the system (if that is the stated purpose)?

Conclusion

We began the paper with a quote from Santayana that emphasizes the importance of knowing our ultimate goal and keeping it clearly in mind as we build the next generation of assessment and accountability systems. Unless we have a clear and common understanding of that goal, it is only by chance that the systems we build will move us closer to attaining it. Unless we have a clear and common understanding of that goal, we can neither measure our progress toward attaining it nor even know when or whether we have attained it. For those reasons, we frame the question of the next generation of assessment and accountability systems in terms of where we want to be in 15 to 20 years. With that goal in mind we can determine the best next steps in improving our current assessment and accountability systems and not merely make incremental steps based on the shiny, new technology or policy.

As we conclude this paper, the thoughts expressed in a second quote from Santayana are equally important:

> *Progress, far from consisting in change, depends on retentiveness. When change is absolute there remains no being to improve and no direction is set for possible improvement: and when experience is not retained, as among savages, infancy is perpetual. Those who cannot remember the past are condemned to repeat it.*

In considering the next generation of assessment and accountability systems to support our reformed public education system, it is important to remember and learn from efforts of the past. We must understand the ways in which our current goals are similar to and different from previous goals, and we must understand the forces that supported our efforts and the barriers that stopped us from meeting those goals. Only with that level of understanding can we hope to accurately determine which aspects of the current system can be maintained as is, which aspects need to be improved, and which need to be discarded and replaced. To be sure, we will not always make the correct decisions based on our interpretation of the past, and changing conditions or goals may limit its relevance.

With that perspective in mind, we close this paper with the words that James Conant used in 1959 to conclude The American High School Today: A First Report to Interested Citizens in which he concluded that reform could occur within the current context of public secondary education[9]. In considering our goals for public education, reflect on the conclusions that we will make about the level of change required in the basic patterns of public education; the importance of a public that is informed, understands, and supports reform efforts; and the belief that reform must occur on a school-by-school and community-by-community basis.

[9] By the time of his second report only eight years later in 1967, Conant was already addressing fundamental changes in public education that we still struggle with today related to the need for more equitable methods of funding public schools, unequal opportunities to learn, and the expansion of the concept of public education beyond high school (Conant, 1967).

> I am convinced American secondary education can be made satisfactory without any radical changes in the basic pattern. This can only be done, however, if the citizens in many localities display sufficient interest in their schools and are willing to support them. The improvements must come school by school and be made with due regard for the nature of the community. Therefore, I conclude by addressing this final word to citizens who are concerned with public education: avoid generalizations, recognize the necessity of diversity, get the facts about your local situation, elect a good school board, and support the efforts of the board to improve the schools. (Conant, 1959, p. 96)

References

AERA, APA, NCME (1999). *Standards for Educational and Psychological Testing*. Prepared by the Joint Committee on Standards for educational and psychological testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Washington, DC: American Educational Research Association.

Baker, E.L. (1988). *Mandated tests: Reform or quality indicator?* CSE Technical Report 283. Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bock, R.D., Mislvey, R., and Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher*, 11, 3 pp. 4-11, 16.

Braun, H. (2008). *Vicissitudes of the validators*, presented at the Reidy Interactive Lecture Series in Portsmouth, New Hampshire, September, 2008.

Claremont School District v. Governor, *138 N.H. 183* (1993), downloaded from http://www.claremontlawsuit.org/Claremont%20I%20web.htm on June 4, 2009.

Claremont School District v. Governor, *142 N.H. 462* (1997), downloaded from http://www.courts.state.nh.us/supreme/opinions/1997/school.htm on June 4, 2009.

Claremont School District v. Governor, *144 N.H. 210* (1999), downloaded from http://www.courts.state.nh.us/supreme/opinions/1999/clarprac.htm on June 4, 2009.

Claremont School District v. Governor, *147 N.H. 499* (2002), downloaded from http://www.courts.state.nh.us/supreme/opinions/2002/0204/clare019.htm on June 4, 2009.

Conant, J.B. (1959). *The American high school today: A first report to interested citizens.* New York: McGraw-Hill.

Conant, J.B. (1967). *The comprehensive high school: A second report to interested citizens.* New York: McGraw-Hill.

Department of the Interior: Bureau of Education (1918). *Cardinal principles of secondary education: A report of the Commission on the Reorganization of Secondary Education*, appointed by the National Education Association, Bulletin, 1918, No. 35. Washington, DC: Government Printing Office. Downloaded on June 4, 2009 from http://www.archive.org/details/cardinalprincipl00natiuoft.

Elmore, R.F. and Rothman, R (Eds.) (1999). *Testing, teaching, and learning: A guide for states and school districts.* Washington, D.C.: National Academy Press.

Forster, M. & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability, in Wilson, M. (ed.) *Towards coherence between classroom assessment and accountability*. Chicago, IL: National Society for the Study of Education.

Linn, R.L. (1995). *Assessment-based reform: Challenges to educational measurement*, William H. Angoff Memorial Lecture presented at Educational Testing Service, Princeton, New Jersey, on November 7, 1994. Princeton, NJ: Educational Testing Service.

Messick, S. (1989). Validity. In R.L. Linn (ed.) *Educational Measurement*, 3rd Edition. New York, NY: Macmillan.

Mislevy, R.J. (1996) Test theory reconceived, in *Journal of Educational Measurement*, 33, 4, pp. 379-416.

Moss, P. (1994). Can there be validity without reliability?, in *Educational Researcher*, 23, 2, pp. 5-12.

National Education Association (1894). *Report of the Committee of Ten on secondary school studies with reports of the conferences arranged by the committee*. New York: American Book Company. Downloaded on June 4, 2009 from http://www.archive.org/details/reportofcomtens00natirich.

Pellegrino, J.W., Chudowsky, N., Glaser, R. (eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Petit, M. (2009). *Vermont Mathematics Partnership Ongoing Assessment Project (OGAP)*. downloaded from http://margepetit.com on June 4, 2009.

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for education reform. In B.R. Gifford & M.C. O'Connor (eds.) *Changing assessments: Alternative views of aptitude, achievement, and instruction*. Boston: Kluwer Academic Publishers.

Rose v. Council for Better Education, 790 S.W.2d 186,60 Ed. Law Rep. 1289 (1989). Downloaded from http://www.wku.edu/library/kera//rose.htm on January 6, 2003.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Francisco: Jossey-Bass Publishers.

Rothstein, R. (2009). *Getting Accountability Right*. Commentary in Education Week, January 28, 2009, Vol. 28, Issue 19, pp. 26, 36.

Santayana, G. (1905). *The Life of Reason: The phases of human progress*. Downloaded as a Project Guteberg ebook from http://www.gutenberg.org/files/15000/15000-h/15000-h.htm on June 4, 2009.

Wilson, M. & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR assessment system, in Wilson, M. (ed.) *Towards coherence between classroom assessment and accountability*. Chicago, IL: National Society for the Study of Education.

Ysseldyke, J. (2008). *Alternative explanations when formative assessment shows that interventions did not work*, presented at the CCSSO National Conference on Student Assessment, Orlando, FL, June 2008.