A Historical Perspective on Validity Arguments for Accountability Testing

CSE Report 654

Edward Haertel
Stanford University

Joan Herman
CRESST/University of California, Los Angeles

June 2005

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Edward Haertel, Stanford University and Joan Herman, CRESST/UCLA, co-Project Directors

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute for Education Sciences, U.S. Department of Education.

A HISTORICAL PERSPECTIVE ON VALIDITY ARGUMENTS FOR ACCOUNTABILITY TESTING

Edward Haertel
Stanford University

Joan Herman
CRESST/University of California, Los Angeles

Using achievement tests to hold students and schools accountable seems an obvious idea. Students come to school to learn. Tests show which students, in which schools, are meeting learning standards and which are not. Those students and schools that are falling short should be held accountable. Of course, the rationales for accountability testing programs are much more complex than that, as are testing's effects, both intended and unintended. In this chapter, we describe various rationales for accountability testing programs over the past century. This history forms the backdrop for current test-driven reforms, including Public Law 107-110, the No Child Left Behind Act of 2001 (NCLB), which was signed into law in January 2002. Our goals are first, to illustrate the diversity of mechanisms whereby testing may affect educational practice and learning outcomes; and second, to show that while many of the same ideas have recurred over time in different forms and guises, accountability testing has become more sophisticated**.** We have a better understanding today than in the past of how to make accountability testing an effective policy tool, although it remains to be seen if we will make the best use of this understanding. The NCLB legislation incorporates various testing policy mechanisms. It relies on testing to focus attention on valued learning outcomes; to spur greater effort on the part of administrators, teachers, and students; to help parents become better informed about school quality; and to direct the allocation of educational resources, including within-school allocations of time and effort, toward groups of students that have lagged behind. Companion federal initiatives rely on testing to identify and to promote effective instructional programs. A look back may offer some insight into both the promise and the pitfalls of contemporary policies.

A theory of action for educational reform typically embodies one or more intended uses or interpretations of test scores. Testing is usually just one part of a

more comprehensive reform strategy. For example, assessments might be expected to identify students requiring remedial assistance, focus attention on teachers whose students are doing especially well or poorly, identify schools where additional resources are needed, or draw attention to achievement disparities among demographic groups. Other elements of the reform strategy would address the delivery of the remedial assistance, reward or sanction deserving teachers, or allocate needed resources. Testing may also be expected to further reform goals in ways less directly tied to the information provided by the scores themselves. For example, a testing program may be expected to clarify learning objectives for teachers or to encourage them to focus on the (tested) basics instead of (untested) "frills," to induce students to work harder, or to focus public attention on issues of school quality and resources. These various mechanisms each imply some *interpretive argument* (Kane, 1992) that could be set forth by way of justification. Interpretive arguments are rationales, often implicit, that might explain exactly how accountability testing is expected to be beneficial. Obtaining and weighing evidence to support or refute the interpretive argument is the business of test validation.

More ominous interpretive arguments may also be formulated. Perhaps accountability testing merely offers the public some hollow assurance as to elected officials' commitment to education. After all, those who propose new testing programs are likely to see achievement gains in 2 or 3 years, right on schedule for the next election (Linn, 2000). Perhaps by reinforcing categories of success and failure, testing contributes to the reproduction of social inequality (Varenne & McDermott, 1999). Testing may subtly shift the blame for school failure from inadequate school resources, poor teacher preparation, or out-of-school factors to teachers and students who are "simply not working hard enough," and thereby divert attention from more costly, more needed reforms.

The next sections of this chapter describe testing during different periods from the early 20th century to the present. Most of these periods are characterized by one or another predominant modes of test use in education. Where appropriate, we refer back to the earlier roots of testing applications characteristic of a given period.

The Turn of the 20th Century: The Birth of Educational Testing

Expectations for educational accountability and student assessment have come a long way since 1864, when the Reverend George Fisher of Greenwich Hospital

School in England put forth, apparently to no avail, the idea of a "Scale-Book," which would:

> contain the numbers assigned to each degree of proficiency in the various subjects of examination: for instance, if it be required to determine the numerical equivalent corresponding to any specimen of 'writing,' a comparison is made with the various standard specimens, which are arranged in this book in order of merit; the highest being represented by the number 1, and the lowest by 5, and the intermediate values by affixing to these numbers the fractions 1/4, 1/2, or 3/4. (cited in Ayers, 1918, p. 9)

The scale book in turn would be used to apply to each student "a fixed standard of estimation that could be used in determining the sum total …or value of any given set of results" (cited in Ayers, 1918, p. 10, where it is attributed to E. B. Chadwick in *The Museum, A Quarterly Magazine of Education, Literature and Science*, Vol. II, 1864). Long too since 1894, when Dr. J. M. Rice in the United States first proposed and was ridiculed for the idea of using an objective standard—in his case a test of 50 spelling words—to compare the relative effectiveness of methods used in different schools (as related by Leonard Ayres, *History and Present Status of Educational Measurements*, 1918, pp. 9-15).

Fisher and Price were forbearers to what Ayers (1918) credits as "the real beginning of the scientific measurement of educational products" (p. 12), the publication of the Thorndike Scale for the measurement of merit in handwriting in March, 1910, and E.L. Thorndike's subsequent persuasion on the necessity for measurement and the need to experiment with tests and scales (see, e.g., Thorndike, 1910, *The Contribution of Psychology to Education*). By 1916, Thorndike and his students had developed additional standardized tests in reading, language, arithmetic, spelling, and drawing (Office of Technology Assessment, 1992). From the beginning, Thorndike pushed concepts that have wide currency today—that education involves the measurement of complex endeavors with endless dimensions from which we must abstract concrete representations for measurement; that scale matters; that the validity of measures must be confirmed with empirical evidence; that reliability and accuracy are essential and require multiple measures—the use of single tasks or items is not sufficient—and that assuring fairness is a challenging concern (see Thorndike, 1918). With regard to the latter, Thorndike particularly advocated the need for and suggested methods "designed to free measurements from certain pernicious disturbing factors, notably unfair preparation for the test,

inequities in interest and efforts, and inequalities in understanding what the task is" (Thorndike, 1918, p. 23).

Early attentive to today's concern with value-added methodologies, Thorndike (1918, p. 16) observed that "education is concerned with the changes in human beings," and its effectiveness could be judged by differences in student behavior—things made, words spoken, acts of performance, etc.—from one point to another. And he predicted the many users who could benefit from such measures—scientists, administrators, teachers, parents, students themselves; and the many uses to which measurement could be put, for example, determining the effects of different methods of teaching or of various features of schools; determining the achievement of total educational enterprises or systems; and even giving individual students information about their own achievement and improvement to serve both motivational and guidance purposes.

During this same period, city school systems, starting in 1911-1912 with New York and moving soon after to Boston, Detroit, and other cities, began to incorporate tests in fledgling efforts to evaluate the results of public schools. It was during this time that educational testing and evaluation gained its first strong foothold. Early tests had a strongly norm-referenced character and were sometimes poorly aligned with learning objectives. But, by the 1930s, power tests had begun to supplant speed tests, and an array of new tests became available to measure basic skills, reasoning, and application of knowledge (Findley, 1963). And, in 1929, E. F. Lindquist, at the University of Iowa, initiated the first statewide testing program, using the Iowa Tests of Basic Skills. These tests were soon made available outside the state of Iowa and added impetus to the shift in testing away from sorting and selecting and back toward diagnosis and remediation (Office of Technology Assessment, 1992, pp. 122-124).

While E. L. Thorndike, at Columbia University, focused primarily on achievement testing, another testing movement was also taking hold, largely led by L. M. Terman, at Stanford University. Terman was among the psychologists who developed the Army Alpha and Beta examinations used to screen and classify recruits after the United States declared war on Germany in 1917. Following the war, Terman and others were eager to apply their new science of mental measurement to the improvement of education. In The Intelligence of School Children, Terman (1919, p. xiv) stated that the Army tests "demonstrated beyond question that the methods of mental measurement are capable of making a contribution of great value to army

efficiency....That their universal use in the schoolroom is necessary to educational efficiency will doubtless soon be accepted as a matter of course." Terman's prediction proved accurate. In 1926, a national survey of urban schools found that over 85% of them used intelligence tests as one basis for classifying children into homogeneous classroom groupings (Chapman, 1979). Tracking students by ability fit well with the emphasis at the time on scientific management in education. And lower intelligence test scores among nonwhites and children of immigrants offered a comfortable explanation for achievement disparities. As students in lower tracks received less rigorous instruction, the predictions from the IQ test scores used for tracking became self-fulfilling prophesies. In his study of the intelligence testing movement in education, Chapman (1979) emphasizes that tracking and other schemes for differentiating the curriculum offered to different children predated the intelligence testing movement. IQ testing did not give rise to tracking, but instead reinforced the practice by providing new classification methods and a stronger scientific rationale. The use of IQ tests for ability grouping persisted into the latter half of the 20th century.

Thus, two fundamental functions of measurement were evident from the beginning of educational testing. One function is sorting and selecting, comparing students to one another for purposes of placement or selection. The second is improving the quality of education. At times, these two categories overlap, as when tests are used both to determine which students merit a high school diploma and to spur greater student effort to meet the standard set. As will be seen, these two broad functions recur again and again.

New Functions for New Forms of Evaluation in the Eight-Year Study

In the 1930s, planning began for the Eight-Year Study, which was to investigate the effect of applying the ideals of progressive education to the high school curriculum. Dr. Ralph Tyler, of the University of Chicago, established a new objectives-based framework for testing and laid out a strong role for assessment in curriculum development and improvement. Formative assessment and continuous improvement models coined decades later have their roots in Tyler's framework. As he later articulated in *Basic Principles of Curriculum and Instruction* (Tyler, 1949), Tyler stressed four principles: Define appropriate objectives; establish useful learning experiences; organize learning experiences to have maximum impact; and evaluate

whether the objectives have been achieved, revising as necessary those aspects of learning that were not effective.

In addition to promoting the use of objectives-based assessments to judge program effectiveness, the Eight Year Study was also significant in recognizing that radical changes could not be made to curriculum and instruction unless influential student evaluation methods were changed at the same time. Students from the 30 progressive high schools involved in the study were evaluated using specially designed "comprehensive evaluations," the term by which Tyler and his colleagues referred to their tests and examinations, and agreement was obtained from over 300 colleges to accept the evidence of these evaluations in lieu of more conventional transcripts and examination results (Madaus, Stufflebeam, & Scriven, 1983; Smith, Tyler, et al., 1942). As participating teachers and schools engaged in curriculum revision and instructional improvement, they also worked with Study staff to develop new measures of their learning goals. These included not only such traditional academic concerns as the application of general science principles, but also scales of beliefs, interest indices, and responses to social problems. Information from these assessments, combined with teachers' observations and judgments, was used to develop comprehensive records of student performance that were to be used by colleges. It is of interest to note that the records were to include *descriptions* and not scores to characterize student accomplishment.

Acknowledging the common purposes of grading students, instructional grouping, and reports to parents, Smith, Tyler, et al. (1942, pp. 7-10) went on to discuss five additional, broader purposes these new evaluations would serve. The first three purposes were checking the effectiveness of educational institutions, checking the effectiveness of specific educational programs or school policies, and providing a basis for sound guidance of individual students. The fourth was "to provide a certain psychological security to the school staff, to the students, and to the parents" (p. 9). Especially in the context of an innovative educational program, these authors viewed a rigorous, comprehensive testing program as important in reassuring the participants that learning objectives were being met. They suggested that without a credible school-based testing program aligned to their progressive learning goals, external examinations like traditional scholarship tests or college entrance examinations might exert an undue influence on teachers' efforts, simply because they could provide some tangible evidence of success or failure. The fifth purpose was "to provide a sound basis for public relations" (p. 10). A strong testing

program would reassure the community served by a school, providing "concrete evidence" of its accomplishments.

As the development of the comprehensive evaluations progressed, another purpose emerged:

> As the evaluation committees carried out their work, it became clear that an evaluation program is also a potent method of continued teacher education. The recurring demand for the formulation and clarification of objectives, the continuing study of the reactions of students in terms of these objectives, and the persistent attempt to relate the results obtained from various sorts of measurement are all means for focusing the interests and efforts of teachers upon the most vital parts of the educational process. (Smith, Tyler, et al., 1942, p. 30)

In summary, the student assessments ("comprehensive evaluations") in the Eight-Year Study served to monitor student progress and guide instructional planning. They also afforded school-level accountability and were used in the evaluation of educational programs and policies. Much like performance assessments in the 1990s, these comprehensive evaluations in the 1930s were also intended to limit the influence of other forms of assessment (traditional scholarship tests or entrance examinations) that were viewed as less progressive and that could otherwise exert undue influence over curriculum and instruction. Again anticipating hopes expressed 50 years later for performance assessments, Tyler's comprehensive evaluations were intended to educate the public about new kinds of learning objectives and to clarify teachers' own understandings of their educational goals.

Measurement-Driven Instruction and Criterion-Referenced Testing

From the 1950s through the 1970s, the principal focus of theory and application in educational testing was measurement-driven instruction, a model which showed strong roots in the Tyler Rationale (e.g., Bloom, Hastings, & Madaus, 1971; Tyler, 1949). This educational testing model found its greatest application at the elementary school level, although curricula were designed along the same lines for learners of all ages, including adults. Material to be taught was analyzed into a series of narrow, carefully sequenced learning objectives (*learning units* or *frames*), each accompanied by a highly focused diagnostic test. These brief, frequent tests were used to guide instruction for individual learners; passing a test was required to proceed to the next unit or frame. The earliest program of this kind was probably the Winnetka Plan, described by one of its developers, Carleton W. Washburne (1925),

in the Twenty-fourth NSSE Yearbook. Under this plan, students worked largely independently, using textbook lessons and mimeographed materials covering narrow learning objectives in a prescribed sequence. Self-tests could be used to monitor progress and to help students determine when they were prepared to take teacher-administered examinations. A record was kept of the date on which a pupil mastered each successive objective.

The 1950s brought a resurgence of interest in such fine-grained task analysis and individualized instructional management. There was considerable optimism at the time that individualized instruction using carefully designed curricula and associated tests could revolutionize schooling practices, as educational psychologists sought to apply principles from the behaviorist psychology of the time (Glaser, 1960). A new science of education was envisioned, with individualized instruction enabling virtually all children to succeed. The psychologist B. F. Skinner, and others, had shown that complex patterns of animal behavior could be shaped incrementally using carefully scheduled reinforcements (Skinner, 1953). Skinner drew implications for human learning from his work with animals, proposing a model for teaching in which the material to be learned was presented in a series of small steps, with probes to check understanding and immediate feedback on correctness. Borrowing from the language of computers, this teaching approach was called "programmed instruction." A critical feature, in the language of the time, was reinforcement of desired behaviors, in this case, correct responses to test questions. Programmed instruction could also be delivered by "teaching machines." Although histories trace development of teaching machines back to "a spelling machine patented in 1866" (Gotkin & McSweeney, 1967, p. 257), major figures in the development of these machines were Sidney L. Pressey, around 1914, and B. F. Skinner, from the 1940s into the 1960s. Early teaching machines were intended as adjuncts to teaching, but Skinner developed the notion of machines that could offer instruction with little human intervention (Gotkin & McSweeney, 1967; Skinner, 1960).

In 1956, Dr. Benjamin S. Bloom and colleagues at the University of Chicago published their *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain* (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). Bloom had been a student of Tyler's at the University of Chicago and shared Tyler's belief that the design of curriculum and instruction must begin with clearly stated objectives. The "Bloom Taxonomy" contributed substantially to the popularity of measurement-driven instructional approaches by showing how test items could be created to measure

"higher-order thinking" (analysis, synthesis, evaluation) as well as "lower-level" learning outcomes like knowledge, comprehension, and application. The taxonomy also gave teachers and curriculum developers a common language to talk about different kinds of learning objectives. Along with Bloom's taxonomy, Robert Mager's (1962, 1975, 1984) *Preparing Instructional Objectives* helped popularize the idea of using tests for fine-grained instructional management, showing teachers how to formulate narrow learning objectives in measurable terms.

Using tests to inform fine-grained instructional decisions entailed a qualitatively different kind of test interpretation from testing for sorting or selection. Instead of interpreting students' scores with reference to the performance of a norm group, as with rankings or percentiles, the score of an individual student was compared to a fixed mastery criterion to determine whether that individual was ready to proceed. The idea of tests designed to show directly what an examinee was able to do, without reference to the performance of anyone else, was formalized by Glaser (1963) as "criterion-referenced testing." In this brief, seminal paper, Glaser articulated ideas that could be traced back to the Reverend George Fisher's proposed scale books. Measurement-driven instruction added the notion of a specific score level denoting "mastery" to this idea of a fixed, criterion-referenced measurement scale (Popham & Husek, 1969).

In the 1960s and 1970s, various models and curricular materials were developed that relied on criterion-referenced testing for individualized instructional management. Perhaps the best known system of this kind was Bloom's (1968) "Mastery Learning" model. Under mastery learning, the material to be taught was divided into a series of units to be mastered sequentially, and mastery tests were created for each unit. End-of-unit tests indicated which students were ready to move on to the next learning unit and which were not. Those not yet demonstrating mastery were re-taught, ideally using approaches different from the initial instruction, including peer tutoring. The goal was to enable virtually all students to attain mastery by assuring that each learner possessed all the prerequisite "cognitive entry behaviors" before embarking on the next unit of instruction. These test-driven instructional systems were a subject of the Sixty-sixth NSSE Yearbook (Lange, 1967). In the Sixty-eighth NSSE Yearbook, Lindvall and Cox (1969) described still other instructional programs designed along these lines, especially the Individually Prescribed Instruction (IPI) Mathematics Project, which they helped to develop at the Pittsburgh Learning Research and Development Center.

Theoretical development of measurement-driven instruction moved beyond the use of post-tests to check students' readiness for the next unit in a fixed sequence. Branching was added to offer supplemental material for students who did not master a unit following the initial presentation. Pretests were introduced to guide the planning of instruction. Glaser and Nitko (1971, pp. 631-632) compared teaching without first making a "detailed diagnosis ... of the initial state of the learner" to "prescribing medication for an illness without first examining the symptoms." The great hope, however, was that distinct kinds of instruction might eventually be offered, tailored to each individual student's own learning aptitudes. This was the vision of "aptitude-treatment interaction" (ATI) research set forth in Lee J. Cronbach's (1957) influential paper on "The Two Disciplines of Scientific Psychology" and later summarized in Cronbach and Snow's *Aptitudes and Instructional Methods* (1977). Rather than forcing all students to adapt to a single mode of instruction and sorting them according to their degrees of success, instruction might instead be adapted to each student's needs, enabling nearly all to attain levels of success previously enjoyed by only a few. To fully realize this vision of adaptive instruction, not only the student's prior knowledge, but also aptitudes were to be assessed, as explained by Glaser and Nitko (1971, pp. 643-645):

> In terms of decisions to be made, the information required is that which answers the question, Given that this student has been located at a particular point in the curriculum sequence, what is the instructional alternative that will best adapt to his individual requirements and thus maximize his attainment of the next instructionally relevant objective? ... It is probably true that a single test of the conventional type now published and used in schools will not be able to provide all the data ... required in an adaptive instructional system.... The basic assumption underlying nonadaptive instruction is that not all pupils can learn a given instructional task to a specified degree of mastery. Adaptive instruction, on the other hand, seeks to design instruction that assures that a given level of mastery is attained by most students.

The aptitude measures required would go far beyond the one-dimensional rankings provided by the IQ tests of decades before. The hope was for tests of specific abilities that could be used to prescribe the optimum form of instruction for each learner. Most scholarship at the time emphasized the need for research on task analysis, individual differences, and alternative modes of instruction. Gagné (1965) showed how complex behaviors could be analyzed into elaborate "learning hierarchies," but it is fair to say that the design of tests and instructional materials featuring "criterion-referenced testing" quickly outpaced the available research. As

"Criterion-Referenced Testing" grew into a popular movement, its central principles were compromised. At the same time, it became clear that while the behaviorist principles of task analysis at the time might be suitable for beginning instruction in reading (e.g., letter-sound correspondences) and mathematical computation, they worked less well for more complex kinds of learning outcomes. And, despite substantial research investments, stable aptitude-treatment interactions proved elusive, with few exceptions. Nonetheless, criterion-referenced interpretations of carefully designed tests have figured in more recent test-based reform initiatives, including performance assessment and standards-based reform.

Educational Testing for Program Evaluation

At the same time as narrow criterion-referenced tests were being used to guide day-to-day classroom instruction, there was increasing use of broader summative tests, often covering a year or more of instructional content, for program evaluation. The Soviet Union's launch of Sputnik in 1957 prompted broad concern over the competitiveness of U.S. students and the quality of U.S. education. One outcome was the development of new curricula in mathematics and the sciences, under the auspices of the National Science Foundation. As new instructional materials like BSCS Biology, PSSC Physics, and CHEM study chemistry came into widespread use, the National Science Foundation and other sponsors began to require evaluations of their effectiveness (Cronbach, et al., 1980). A review of findings from over 20 such evaluations found that "groups studying from the innovative curriculum scored higher on virtually every test which favored the innovative curriculum, *but* groups studying from the *traditional* curriculum scored higher on a substantial number of comparisons in which the test content more nearly resembled what *they* had studied" (Walker & Schaffarzick, 1974, p. 91, italics in original). Thus, the curricula that fared best in these evaluation studies were those that included new tests specifically aligned to new learning objectives.

Passage of the Elementary and Secondary Education Act of 1965 (ESEA), part of President Lyndon Johnson's "War on Poverty," greatly expanded the use of formal evaluations of educational programs. Under Title I of the ESEA, school districts received federal funds to provide extra academic support for children from low-income families. Extensive regulations were put in place to help assure that the money was spent appropriately. In addition, at Sen. Robert Kennedy's insistence, an annual testing requirement was added for all children in Title I programs, to

determine whether the programs were meeting their objectives (Cross, 2003). The idea of evaluation was not new, but the mid-'60s brought federally funded educational evaluations of unprecedented size. This use of evaluation, in particular of objective test data, for program oversight fit well with the rational management practices pioneered in the military under the direction of Robert McNamara and then applied more widely under the Johnson administration (Lagemann, 1997).

Proponents of compensatory education hoped that evaluations documenting program effectiveness would build support for these social programs, but results were disappointing. Evaluations of ESEA, Head Start (aimed to pre-school students), and Follow Through (providing continuing support to Head Start students as they progressed through school) failed to find evidence that the programs were effective. Madaus, Stufflebeam, and Scriven (1983) observed that one problem with these evaluations was the sorts of tests they employed. The standardized tests available at the time were designed and normed to provide accurate individual measurements and stable rankings of children of average ability for the grade tested. They measured broad abilities developed over years of schooling. As general ability tests, they were insensitive to short-term instructional effects. Also, they were sometimes too difficult for disadvantaged students, and were not aligned with learning objectives appropriate for Title I student populations.

Another major, federally commissioned study of the period was the Equality of Educational Opportunity Survey (EEOS), led by James Coleman (Coleman, et al., 1966). This study was expected to demonstrate that profound achievement differences associated with social class and race were attributable to disparities in educational resources. Instead, the authors concluded that the quality of schooling had little effect independent of a child's family background and out-of-school environment. Rather than supporting increased outlays to redress inequalities in educational opportunity, the EEOS appeared to show that changes in school quality would likely have little effect.

Minimum Competency Testing

For many reasons, the 1970s brought growing discontent with public education. The apparent failure of compensatory education and the seeming intractability of achievement gaps was one contributing factor. Another was a widely publicized, pervasive test score decline that began in the late 1960s and continued through the 1970s. Also, the problem of high levels of youth

unemployment received much attention (e.g., National Research Council, 1983), and inadequate academic skills were viewed as contributing to the problem (Resnick, 1980). The media and alarmist reports from education reform panels fueled a popular perception that pupils were just being passed along from grade to grade, and a high school diploma no longer meant much of anything (Office of Technology Assessment, 1992). In response to disillusionment with educational reform policies focused on "inputs" (better resources, better curricula, new teaching methods), policymakers shifted attention to interventions that focused on outcomes. There was some experimentation with performance contracting (monetary incentives for teachers whose students reached specified benchmarks) and with accountability systems that tied state funding to school-level test scores, but these were short-lived (Cohen & Haney, 1980). The approach that caught on was minimum competency testing, an outgrowth of the "back to basics" movement of the 1970s. The minimum competency test (MCT) was a basic-skills test, usually in reading and mathematics. Typically, students were required to pass the MCT in order to receive a regular high school diploma, although MCTs could be used in other ways. Interpretation was criterion-referenced. That is, an absolute level of performance was required, represented by a passing score that was locally determined. The actual level of proficiency required was probably around the eighth-grade level or lower in most cases. Indeed, some found that MCT reforms tended to be largely symbolic, in that proficiency levels were lowered so that politically unacceptable numbers of students would not fail, and the standard thus became so diluted that few systematic changes in instruction or learning occurred (Ellwein & Glass, 1986). Nonetheless, a national study found that students who did not pass their tests on the first try were more likely to drop out of school (Catterall, 1989).

Minimum competency testing began in a few school districts as early as 1962. By 1980, statewide minimum competency testing requirements had been implemented in 29 states, most having been initiated in 1975 or later. (In some states, the tests were used to place students in remedial programs or were required for grade promotion as opposed to high school graduation.) By 1985, although 33 states required students to take an MCT, only 11 still made passing the test a requirement for the high school diploma. Passing rates on MCTs in many states rose rapidly from year to year (Popham, Cruse, Rankin, Sandifer, & Williams, 1985). Despite these gains, and positive trends on examinations like the National Assessment of Educational Progress (NAEP), there is little evidence that MCTs were the reason for

improvements on other examinations; and the improvements in passing rates on MCTs themselves may have reflected little more than the effects of drill and practice narrowly focused on tested skills, and possibly also the effect of increased dropout rates (Catterall, 1989; Shepard, 1991b). Over time, popular concern shifted from an emphasis on "basic skills" toward complex, "higher order thinking" skills, and the MCT movement faded (Office of Technology Assessment, 1992, Ch. 2).

Performance Assessment and the Growth of the Standards Movement

As the 1980s dawned, the United States already was replete with standardized tests and data on student achievement. Results from annual tests stimulated by the 1965 Elementary and Secondary Act and its reauthorizations were routinely published in local newspapers; minimum competency testing was in full swing across the country; the federally funded National Assessment of Educational Progress (NAEP, now known as "The Nation's Report Card") had been reporting periodically on the performance of 9-, 13-, and 17-year-old students in reading, mathematics, writing, science, and additional content areas since its inception in 1969[1]; the College Board annually reported results of its Scholastic Aptitude Test (SAT); and international assessments were growing in popularity. However, as U.S. policymakers and the public looked at the data, they did not like what they saw. American business, furthermore, was concerned about international competitiveness and dissatisfied with the preparation of the entering workforce.

A Nation at Risk

Focusing particularly on NAEP, SAT, and international comparisons, a prominent national commission declared the country "A Nation at Risk" (National Commission on Excellence in Education [NCEE], 1983). At a time when the scientific and technical demands of the workforce and the citizenship were growing, the Commission found student achievement in decline and urged that students be engaged in the rigorous curriculum they would need for future success. Among its five major recommendations, the Commission advocated that "schools…adopt more rigorous and measurable standards, and higher expectations, for academic performance and student conduct" (NCEE, 1983, p. 27). Making clear a strong public commitment to the twin goals of excellence and equity, to high expectations and to

[1] Since the mid-1980s, NAEP has reported results for Grades 4, 8, and 12, rather than for student cohorts defined by chronological age.

developing all individuals to their highest potential, the Commission specifically recommended that:

> 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…[in order to]: (a) certify the student's credentials; (b) identify the need for remedial intervention; and (c) identify 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. (NCEE, 1983, p. 28)

Presaging today's standards-based accountability systems, the Commission viewed standards and assessment as a major part of the solution in stemming the "rising tide of mediocrity" in American education (NCEE, 1983, p. 5) and gave rise to a number of complementary initiatives in the late 1980s and early 1990s. For example, President George H. W. Bush in 1989 convened the 50 governors in an Education Summit, resulting in agreement on six broad goals that American students should reach by the year 2000, as later reified in the creation of the federally funded National Education Goals Panel (National Education Goals Panel [NEGP], 1991). Two of the goals addressed expectations for student achievement and directly stimulated the development of standards and assessments. Goal 3 specified that:

> American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter, including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy. (NEGP, 1991, p. 10)

Goal 4, in turn, insisted that American students become "first in the world in science and mathematics achievement" (NEGP, 1991, p. 16).

Moving to agree on what such competencies and achievement ought to be, national subject matter organizations, starting with the National Council of Teachers of Mathematics (1989), set about to define subject matter standards in their respective disciplines. The U.S. Secretary of Labor in 1990 appointed the Secretary's Commission on Achieving Necessary Skills (SCANS) to articulate the skills that students need for success in work, and in that same year, the Learning Research and Development Center, led by Lauren Resnick, and the National Center on Education and the Economy, led by Marc Tucker, established the New Standards Project, to

create a voluntary system of academic performance standards and assessments. Under President Clinton, in turn, the Goals 2000: Educate America Act (PL 103-227, 1994) established an initial framework and funding stream to support states and national entities to identify challenging academic content standards, develop measures of student progress, and to link state and local reform efforts to enable students to meet the standards. The 1994 reauthorization of Title I, the Improving America's Schools Act (IASA), built on the Goals 2000 framework and required states to develop or adopt challenging content and performance standards that were to apply to all students, to develop assessments aligned with those standards, and to be accountable for student performance. Attention to inputs in Title I evaluation had waned; schools were to be accountable for performance outputs. Moreover, responding to growing interest in performance assessment, IASA also required that state tests address complex thinking skills and include multiple measures.

Assessment as a Reform Strategy

Content and performance standards represented a watershed in thinking about the role of assessment in reforming and improving schools. The problem with past approaches was becoming clear. With attention to traditional standardized, multiple-choice test results on an upswing in the 1980s, so too was time devoted to testing and test preparation (Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Kellaghan & Madaus, 1991; Smith & Rottenberg, 1991). The net effect was a narrowing of the curriculum to the basic skills that were assessed, a neglect both of complex thinking skills and of subject areas that were not assessed, and a tendency for teachers to mimic the tests' multiple-choice formats in their classroom curriculum (Haertel & Calfee, 1983). Critics saw existing testing—and thus prior iterations of measurement-driven instruction—as driving teaching and learning in the wrong direction, promoting outmoded behaviorist pedagogy that was unlikely to prepare students for success (Haertel, 1999; Herman, 1997; Herman & Golan, 1993; Shepard, 1991a; Resnick & Resnick, 1992). In the emerging view, if assessments were aligned with comprehensive content standards and if expected levels of attainment were codified in ambitious performance standards, then high-stakes testing could be transformed into a positive instrument of educational reform. The problem was not with testing, but was instead with using the wrong sorts of tests.

The New Standards Project and the performance assessment movement attempted to turn the negative into a positive. Cognizant of testing's role in communicating expectations, advocates sought to use performance assessment to

broadcast a new vision of education and to promote a "thinking curriculum" (Gong & Reidy, 1996; Resnick and Resnick, 1992). If what you get is what you assess (WYGWYA, as the expression went then), then assessment needed to reflect the kind of teaching and learning activities which newer views of learning supported and be "tests worth teaching to" (Resnick & Resnick, 1992). Be they open-ended items that asked students to compose or explain their answers; inquiry-oriented research papers, experiments, or demonstrations; essays or artistic expressions; portfolios of varied work over the course of a semester or year; or any number of other options, the essence of performance assessments was that they asked students to *create* something of meaning, were intended to invoke authentic and real world applications, to tap complex thinking and/or problem solving, and were *not* multiple choice. Furthermore, because students typically were asked to construct unique answers, performance assessments usually required substantially more time than their multiple-choice predecessors and needed to be scored by humans, exercising judgment, rather than by machines scanning graphite marks (Herman, Aschbacher, & Winters, 1992; Wiggins, 1992).

Validity and the Meaning of Assessment Quality

The performance assessment movement thus moved to create new standards for student learning by articulating what students ought to be able to do and framed these expectations in the context of *A Nation at Risk*'s and the National Education Goals Panel's calls for rigorous curriculum. In so doing, the performance assessment movement highlighted some limitations of then current conceptions of test quality and created new challenges for educational measurement. Building on Samuel Messick's (1989) thinking on the central role of test use and consequences in validity research, Robert Linn and Eva Baker (1996) argued for an expanded vision of test validity that considered multiple internal and external criteria. Among the internal validity criteria, they suggested were content quality, curricular importance, content coverage, cognitive complexity, linguistic appropriateness, ancillary skills, and meaningfulness of tasks for students; and among the external validity criteria were consequences for students and teachers, fairness, transfer and generalizability, comparability, and instructional sensitivity.

Research showing the fragility of student performance across ostensibly similar performance assessment tasks created special challenges for the generalizability and comparability criteria. That a student did well on one mathematics problem-solving task, for example, did not mean the student would do well on a second such task.

This variability meant that it would take a number of tasks to get an accurate estimate of student achievement in a particular knowledge domain. Then available research, for example, suggested that from 5 to 20 tasks were needed to obtain reliable individual estimates (Baker, 1994; Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993), depending on the knowledge domain and the breadth of its specification. Even at the lower end, given the time requirements of typical performance assessment tasks, the time demand challenged their feasibility for large-scale use. For example, the hours of testing time required to get adequate reliability on the free response sections of 21 Advanced Placement tests ranged from 1 hour 15 minutes (in physics) to 13 hours (in European History) (Linn & Baker, 1996, p. 97). Performance assessment, in short, brought tensions between traditional psychometric views of test quality and broader conceptions focused on consequences.

Challenges in Fairness and Equity

Performance assessment brought new tensions in assessment equity as well. Advocates believed that traditionally low-performing students would likely be more engaged and, thus, be better able to show what they knew on the authentic tasks of performance assessment and hoped that publicly articulating standards would raise expectations and promote richer educational opportunities for economically poor and culturally diverse students. As Baron and Wolf (1996) observed, performance assessment was the tool of choice for those from two distinct perspectives on American education. One view, emanating from *A Nation at Risk*, regarded American education as failing and promoted rigorous assessments as a means to establish consequences for individuals and schools, ultimately serving sorting functions. The second represented the vision of the egalitarian common school, where performance assessments were intended to make public higher expectations for students and to promote dialogue and action about who had the opportunity to attain those standards. At the same time, Gordon and Bonilla-Bowman (1996) questioned performance assessment's ability to bridge these two paradigms and affect the status quo, concerned that historically, such innovations had been "redesigned in ways that continue to exclude those who have been historically underserved" (p. 35). They further raised issues of the adequacy of students' opportunity to learn; the potential for teacher bias in scoring performance assessments; and the possibility for conflicts between students' internal standards, which were likely to be culturally based and unique to each individual, and external

standards, which were likely to be more technical and to reflect the power relationships of the dominant culture. They also noted possible collisions when assessment systems reflecting high standards met the reality of inequities in schooling. As they observed:

> The choice cannot be between denial of opportunity [e.g., a diploma] or acceptance of lower standards. Failure to hold students of color to a common standard because of an acknowledgment of the inferior quality of their previous schooling is crippling, though perhaps not as destructive as exclusion from the opportunity for correction because of the failure to meet arbitrary standards. Obviously, the solution of this problem lies in the direction of more appropriate pedagogical intervention. (pp. 46-47)

Performance Assessment and the Improvement of Teaching

Others too saw great value in performance assessment for promoting pedagogical improvement, echoing some of the thoughts of Tyler and his colleagues 60 years earlier. Beyond communicating higher expectations and modeling meaningful instructional activities, advocates noted the professional development benefits of involving teachers in developing and scoring performance assessments (Goldberg & Rosewell, 2000; Resnick & Resnick, 1992), and the special value of involving teachers in learning communities where, in the context of appraising student work, they could develop common expectations, derive insights on students' thinking and understanding, share successful strategies for improving students' learning opportunities, and get support for changing teaching practices (Darling-Hammond & Ancess, 1996; LeMahieu & Eresh, 1996). For, indeed, performance assessment represented a radical change in the epistemology for teaching. Echoing the hopes for Mastery Learning and similar measurement-driven instructional approaches a generation before, performance assessment was tied to new, developmental views of learning in which support and effort could enable most students to attain high expectations. Proponents hoped it would bring broad acceptance of new definitions of what it meant to know and understand (Baron and Wolf, 1996; Wolf & Reardon, 1996). Grounded in constructivism, new pedagogical theory brought new ideas about the role of assessment in and for learning (Stiggins, 2002) and increasing attention to students' self-assessment, ideas which also ground today's interest in classroom and formative assessment (see Shepard, 2000; and section below).

The Current Accountability Context

Today's accountability and assessment context features many of these same ideas about the role of standards and assessment in improving student learning, ideas advanced by the No Child Left Behind (PL 107-110, 2001; NCLB) mandates. NCLB strengthened the accountability requirements of the 1994 reauthorization by insisting that states implement statewide accountability systems covering *all* public schools and students based on challenging academic content standards in reading and mathematics; annually assess all students in Grades 3-8, plus one high school grade, relative to established standards;[2] and create annual statewide performance targets for schools to assure that all students reach proficiency in both subjects by the year 2014. In the interests of encouraging schools to address the needs of all their students and of reducing the achievement gap, NCLB requires not only that all schools meet their states' annual performance targets in terms of the percentage of students scoring "proficient"—deemed adequate yearly progress (AYP)—but that schools meet them for every numerically significant subgroup at the school, as defined by race or ethnicity, language status, poverty, and disability status. Schools and school districts that fail to meet their achievement targets over time are designated "in need of improvement," and are subject to corrective action leading up to restructuring or reconstitution, all measures aimed at getting them back on track relative to AYP targets. Parents at such schools are given options for sending their children to other schools or for special supplementary services, funded by their local school district. Ironically, while NCLB attempts to focus on the rigorous academic standards that were the hallmark of performance assessment reforms in the previous decade, the cost and feasibility of annual state testing at so many grade levels has probably discouraged the use of extended performance assessments (General Accounting Office, 2003). Even in the face of NCLB requirements for multiple measures and for a range of evidence to support the validity of state tests, state assessments have retreated to the predominant use of multiple-choice items, once again raising the worries of the last decade about attention to the complex thinking and problem-solving skills that underlie truly rigorous academic standards. According to the annual "Quality Counts" report by *Education Week*, in 2004-'05, virtually all states employed multiple-choice tests, and all but four (five with the District of Columbia) used extended response questions in English/Language Arts. Roughly two-thirds of the states employed short-answer questions, but only one

[2] Science standards and assessment must be implemented at selected grade levels by the year 2007.

used portfolios as part of its state accountability system, and performance assessments were not even listed as a category (R. A. Skinner, 2005, p. 87).

The NCLB Theory of Action

Nonetheless, NCLB firmly continues the policy assumption that being explicit about standards for student performance and measuring student progress toward them, coupled with sanctions and incentives, will leverage the improvement of student learning. The intent is to provide a technical assessment system that cannot only measure performance and provide feedback to support improvement, but perhaps more importantly, can serve motivational and symbolic purposes in: establishing the target for reform efforts; communicating to educators, administrators, and parents what is expected; providing incentives and/or sanctions; and thereby stimulating all levels of the education system to focus on achieving the NCLB goals for AYP, ostensibly assuring that all children will be proficient by the year 2014.

Figure 1 shows one view of how accountability is supposed to work, focusing particularly on the quality of classroom teaching and learning necessary to enable students to reach intended standards (Herman, 2004). While the full and coordinated support of all levels and resources of the educational system may be needed to achieve policy goals, it seems axiomatic that students cannot be expected to become proficient unless and until the content and process of their classroom instruction well prepares them to do so. As the figure shows, standards are the basis for accountability assessments and likewise are the targets of classroom teaching and learning. Feedback from the assessment is used to improve learning opportunities for students and to increase their attainment of standards. Because every subgroup of students within the school must attain the adequate yearly progress targets, the intent is to assure that schools and teachers will hold high expectations and provide appropriate opportunities to learn to *all* students, including whatever augmented programs and special services traditionally low-achieving students may need to attain success. Although Figure 1 focuses on the impact of assessments at the classroom level, schools and districts also are expected to use the feedback from state assessments to gauge their strengths and weaknesses, to identify students who may need special help, and to be strategic in taking action and coordinating available resources to improve student performance.

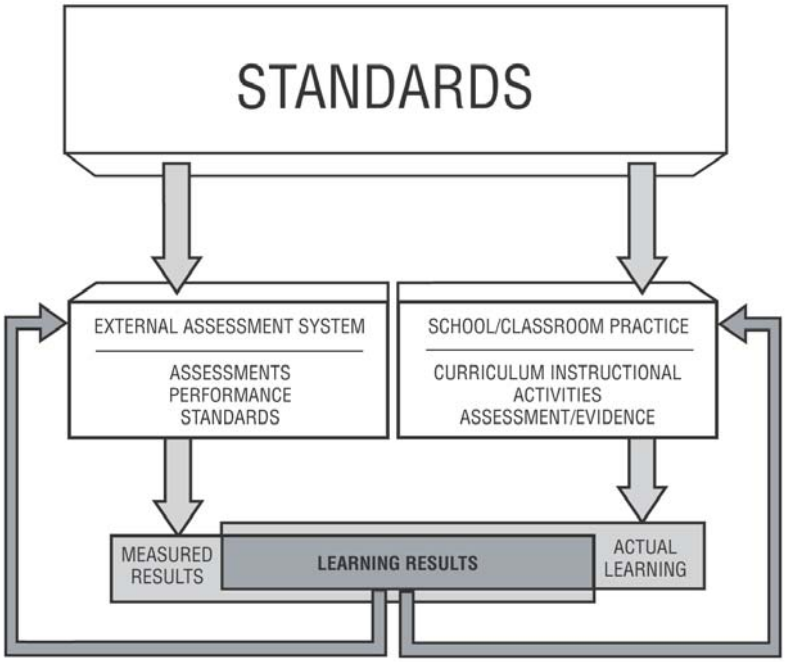

Figure 1. An ideal model for standards-based accountability.

What should be apparent from Figure 1 is the importance of several technical features of the system. First, the alignment of standards, assessments, and classroom instruction is critical to the validity of the system. It is only when the contents and processes of teaching and learning correspond to the standards that students indeed have the opportunity to learn what they need to be successful. And it is only when assessment is aligned with both standards and classroom instruction that assessment results can provide sound information about both how well students are doing and how well schools and classroom teaching are doing in helping students to attain the standards. Because of the centrality of alignment to current policy logic, researchers over the last decade have worked to develop methodologies for assessing it, looking particularly at the match between content and cognitive demand (see, e.g., Porter, 2002; Rothman, et al., 2002; Webb, 1997, 2002). Their work shows an uneven match between standards and state assessments.

Yet even with tight alignment, Figure 1 tries to make it clear that all tests are fallible and can only measure a part of what students are learning. Tests can only assess that which can be measured in the finite time allotted for testing and through

the particular formats employed in the tests—meaning that it is impossible for tests to assess everything that is important. Furthermore, all measures also contain error and thus provide only an imperfect *estimate* of student performance. With the advent of standards-based tests, these imperfect estimates must then be converted into proficiency classifications—based on one of a number of standard-setting methods that have been developed over the last four decades, a process that brings significant technical challenges (Haertel & Lorié, 2004). Based on AYP requirements, percentages of students scoring at or above proficient must be determined for each numerically significant subgroup within a school or district, raising thorny questions about the minimum size needed to achieve sufficient stability of subgroup estimates and creating tensions between technical concerns about the quality of the data and consequences relative to assuring attention to subgroup needs.

Figure 1 also attempts to make clear that state assessments are not the only assessments of importance in the system. The continuous improvement model that accountability envisions means that educators must keep their eyes on student learning, conduct regular assessments *of* and *for* student learning to see how students are doing relative to standards, use the information to understand what students need, and take appropriate, meaningful action based on learning evidence—just as Tyler and colleagues envisioned seven decades ago. We expand later in this chapter on current theory about formative classroom assessment *for* learning (Black, et al., 2004; Stiggins, 2002).

In Figure 1, standards guide curriculum and instruction as well as assessment and all students are given access to the entire content of the agreed upon standards. This may represent an optimal view of standards-based reform, but Figure 2 may better represent the current reality. Research cited earlier strongly suggests that educators, particularly in schools that are under the greatest pressure to show improvement, are teaching to the test, not the standards (Stecher, et al., 1998; Stecher, et al., 2000; Pedulla, et al., 2003). Accountability tests, thus, are the lens through which the standards are interpreted and serve to define the standards. Standards in subjects not tested and standards that are not included in subject matter tests seem to get at most weak treatment in classroom teaching and learning.

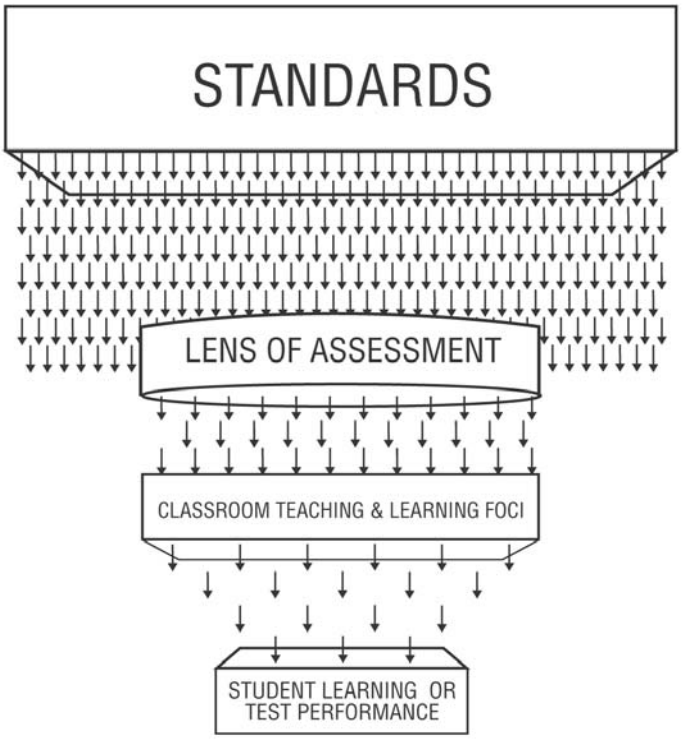

Figure 2. Potential effect of accountability testing on student learning.

As Figure 2 highlights, a focus on tests rather than standards has serious consequences for students. Rather than being exposed to the full breadth of knowledge and skills that society has determined are important for future success, students have the opportunity to learn only a relatively narrow, test-based curriculum. Traditionally low-performing students—the economically disadvantaged, language minority, students of color, and students with disabilities—are most likely to be negatively affected, since their instruction is most likely to focus intensely on reaching proficiency based on state assessment results (see Darling-Hammond, this volume). There is danger of a dual curriculum evolving for these, versus more advantaged students, and of serious equity problems.

Moreover, with the specifics of the test—rather than the essentials of the discipline or meaningful learning—as a primary focus, there also is growing danger of test score inflation. Students may be learning only what is tested, and increases in test scores may not generalize to other situations. Potential mismatches between

tests and standards can lead educators and policymakers to misinterpret test results and fail to address genuine needs.

These same issues and basic theories of action are relevant to high school exit examinations that are growing in popularity across the country. Harkening back to the minimum competency testing reforms of the 1970s and 1980s, and responding to similar concerns from U.S. business and universities about students' preparation, the idea is that high school exit exams will reflect the standards that students must reach for future success. There is an expectation that students will be motivated to learn this material and schools and teachers will be motivated to teach it if passing the test is required for high school graduation. By the year 2003, 19 states required exit examinations and five were scheduled to phase them in over the next few years (Center on Education Policy, 2003, p. 5). Echoing as well the concerns around earlier MCT, the business community worries that current tests are not sufficiently rigorous to assure adequate student preparation (Achieve, 2004). At the same time, there is pushback from parents and communities when unacceptable numbers of students fail to pass the tests. Similarly, the tests raise serious equity concerns, given the disproportionality of passing rates for economically poor students and students of color, and data suggesting a relationship between high school proficiency requirements and students dropping out. Exit examinations may reduce graduation rates, especially for African American and Latino students, English Language Learners, and students with disabilities (Darling-Hammond, et al., 2005).

Formative Assessment for Student Learning

Ultimately, students' needs must be addressed in the teaching and learning opportunities provided for them in classrooms, schools, and/or other settings and, as noted above, any reasonable theory of action linking assessment with student learning would hold an important place for classroom assessment. Just as Ralph Tyler (1949) advocated for the use of assessment to improve curriculum and Benjamin Bloom (1968) argued for the central role of ongoing assessment in mastery learning, so too is feedback on student learning seen as essential in today's classroom teaching and learning (Sadler, 1989). Indeed, Black and Wiliam's (1998a, 1998b) historic meta-analysis found that formative assessment—the use of assessment to provide feedback to teachers and students to modify instruction and enhance learning—produces significant student learning gains beyond those of other available interventions and that it helps to narrow the achievement gap

between low and high achievers. Their work presents a landmark in burgeoning interest in classroom and formative assessment as a strategy for improving student learning.

Wiliam (2004) notes that assessments for accountability and assessments for learning differ not so much in source—external versus classroom—but more importantly in design impetus and use. Assessments designed to serve learning purposes provide feedback for students and teachers that can be used to modify the teaching and learning activities in which they are engaged; such assessment become formative when the feedback is actually used to improve learning (Wiliam, 2004). In terms of designing assessments for learning, the National Research Council (NRC) report, *Knowing What Students Know* (NRC, 2001) insists that such assessments start with a model of student cognition, with assessment tasks designed and interpreted to inform inferences about students' learning, and appropriate next steps, and that assessment design be consistent with modern views of effective pedagogy.

In modern conceptions, moreover, assessment becomes part and parcel of the teaching and learning process. Contemporary cognitive psychology recognizes that knowledge is always actively constructed by learners (NRC, 2000, 2001). A situative perspective reminds us that knowing is a verb before it is a noun—what is acquired through schooling is a set of capabilities for meaningful participation in activity structures; all knowing has a social component (Gipps, 1999). In earlier behaviorist models, assessment served to monitor students' status with regard to relatively static learning goals, so that someone or something could make adjustments as instruction was imparted. In modern conceptions, assessment provides opportunities for students to display their thinking and to engage with feedback that can help them to extend, refine, and deepen their understandings and reach more sophisticated levels of expertise. Portfolio assessments present one example of such assessments, in which students may engage in various types of writing to explore personal or content ideas, may be encouraged to include drafts to show their writing process, and may reflect on their own performance and progress (Calfee & Perfumo, 1996). Teachers' informal questioning during the course of class discussions presents another example, where questions are used to elicit students' understandings and alternate conceptions. Feedback is used to encourage students to confront their misconceptions and to move to higher levels of understanding (Gitomer & Duschl, 1995).

Conclusion

The past century's history of educational testing in the United States shows the varied ways in which assessment has been expected to support educational quality and the improvement of students learning. This chapter has described widespread, popular movements as well as a few pivotal studies that have shaped popular perceptions of education and influenced educational policy. The story told is scarcely one of steady progress toward some inevitable ideal. Again and again, testing movements that seemed unstoppable have fallen short of expectations and faded away, only to be replaced by some new approach. Nonetheless, each new testing initiative has left its traces, and the theories, hopes, and expectations of the past reverberate in present testing policies and practices.

The stated goal of today's standards-based accountability is to help all children reach the same ambitious academic content standards. This is in striking contrast to the use of IQ tests during the first half of the 20th century, which were used to determine which children should be provided an academically rigorous curriculum and which would find such a curriculum beyond their abilities. Aptitude testing, the use of cognitive tests of different kinds to determine students' capabilities and limitations, evolved from the broad use of IQ tests for sorting students along a continuum toward testing of narrower abilities to predict the forms of instruction best suited to individual learners. Today, while some tests of "learning styles" are still in use, testing of individual abilities is largely limited to diagnosis and placement of children with specific cognitive disabilities.

Even though routine IQ testing of all children has largely disappeared, the use of tests for sorting and selecting has continued. Today, however, the sorting function relies on achievement tests rather than aptitude tests. Even the redoubtable Scholastic Aptitude Test, first designed to identify native talent wherever it might be found, has dropped "Aptitude" as its middle name and is instead described as a broad test of achievement and of reasoning *skills* that can be learned and practiced. Minimum competency tests and, more recently, high school exit examinations sort examinees into broad categories of passing and failing, although the stated goal of these tests is to assure that all meet the standard, not to penalize those who fall short. This shift from aptitude toward achievement testing is consistent with the aspiration that all students master a rigorous academic curriculum. Unlike some important learning aptitudes, achievement can be increased through individual effort.

From Washburne's Winnetka Plan through the periods of aptitude-treatment interaction research, teaching machines, and criterion-referenced testing, testing was used to guide instructional decision making for individual learners. Fine-grained tests closely tied to narrow learning objectives dictated the pacing of instruction and provided feedback to students and their teachers. These methods were bound up with a conception of curriculum and instruction that has grown less popular, although some highly scripted basic-skills curricula are still widely used and have strong adherents.

As the goals of education have shifted away from acquiring factual knowledge toward "higher order thinking," the limitations of narrow paper-and-pencil measures have become clear. Although multiple-choice tests are still used to measure many valued learning outcomes, it is recognized today that they are poor measures of some of the most important goals of education. For a short time, it appeared that portfolio-based assessments and performance assessments might hold the key to rapid instructional improvement and the closing of historic achievement gaps that have separated underserved from more advantaged learners. Echoing the hopes of Tyler's Eight-Year Study for new forms of tests to support new forms of learning, performance assessments were adopted uncritically in the 1990s as a tool of educational reform. Disillusionment with high costs, low reliability, and poor student performance on these examinations, coupled with the dramatic increase in amount of testing required under NCLB, have brought a shift back to heavy reliance on multiple-choice tests; the promise and potential of performance assessment remain unfulfilled.

One constant in this changing picture has been the idea that education should produce measurable results. Again and again, policymakers have advanced accountability testing as a means for improving education, each generation responding to the failings of the previous. From the days of Joseph Rice and the school testing programs of the early 1900s, through the Head Start program evaluations of the 1960s and up to the increasingly prescriptive testing requirements of successive ESEA reauthorizations culminating in the No Child Left Behind Act of 2001, policymakers have used tests in an attempt to discover which schools and districts are fulfilling their responsibilities and which are falling short. NCLB responds to the perceived failings of previous accountability testing programs in various ways. Because improvement seemed to come too slowly under state accountability systems, "Adequate Yearly Progress" has been defined in a way

intended to bring all children to "proficient" by 2014. Because achievement gaps have persisted, student subgroups must be tracked separately and each must meet the same annual measurable objectives for a school to demonstrate AYP. Because it matters what is tested, states must use tests aligned with rigorous academic content standards. Because multiple-choice tests alone are poor measures of complex learning outcomes, NCLB calls for the use of multiple measures. These are promising innovations, but history shows that testing alone, in itself, is unlikely to bring about major educational improvement. It remains to be seen how effective today's accountability testing will prove to be in supporting genuine, comprehensive educational reform.

References

Achieve, Inc. (2004). *The expectations gap: A 50-state review of high school graduation requirements*. Washington, DC: Achieve, Inc. (www.achieve.org)

Ayres, L.P. (1918). History and present status of educational measurements. In G. M. Whipple (Ed.), *The measurement of educational products* (Seventeenth yearbook of the National Society for the Study of Education, Part II, pp. 9-15). Bloomington, Ill: Public School Publishing Company.

Baker, E. L. (1994). Researchers and assessment policy development–A cautionary tale. *American Journal of Education*, *102*, 450-477.

Baron, J. N., & Wolf, D. P. (1996). Editor's preface. In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. ix-xiii). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Black, P., Harrison, C., Lee, C., Bethan, M., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8-21.

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

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 46-48.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2), 1-12. (Available from the ERIC Document Reproduction Service, No. ED 053419.)

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals, handbook I: Cognitive domain*. New York: Longmans, Green.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Calfee, R., & Perfumo, P. (Eds.). (1996). *Writing portfolios in the classroom: Policy and practice, promise and peril*. Mahwah, NJ: Erlbaum.

Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for high school graduation. *American Journal of Education*, *98*, 1-34.

Center on Education Policy. (2003). *State high school exit exams: Put to the test.* Washington, DC: Author.

Chapman, D. P. (1979). *Schools as sorters: Lewis M. Terman and the intelligence testing movement, 1890-1930*. Unpublished doctoral dissertation, Stanford University.

Cohen, D. K., & Haney, W. (1980). Minimums, competency testing, and social policy. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing* (pp. 5-22). Berkeley: McCutchan.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex Publishing.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Halsted Press.

Cross, C. (2003). *Political education: National policy comes of age*. New York: Teachers College Press.

Darling-Hammond, L., & Ancess, J. (1996). Authentic assessment and school development. In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 52-83). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Darling-Hammond, L., Rustique-Forrester, E., & Pecheone, R. (2005). *Multiple measures approaches to high school graduation.* Palo Alto, CA: School Redesign Network at Stanford University.

Dorr-Bremme, D., & Herman, J. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph Series in Evaluation No. 11). Los Angeles: University of California, Center for the Study of Evaluation.

Dunbar, S., Koretz, D., and Hoover, H. D. (1991). Quality control in the use of performance assessment. *Applied Measurement in Education*, *4*, 289-303.

Ellwein, M. C., & Glass, G. V. (1986). *Standards of competence: A multi-site case study of school reform* (CSE Tech. Rep. No. 263). Los Angeles: University of California, Center for the Study of Evaluation. (Available from the ERIC Document Reproduction Service, No. ED293883.)

Findley, W. G. (1963). Purposes of school testing programs and their efficient development. In W. G. Findley (Ed.), *The impact and improvement of school testing programs* (Sixty-second yearbook of the National Society for the Study of Education, Part II, pp. 1-27). Chicago: The National Society for the Study of Education, distributed by the University of Chicago Press.

Gagné, R. M. (1965). *The conditions of learning*. New York: Rinehart & Winston.

General Accounting Office. (2003, May). *Characteristics of tests may influence expenses; information sharing may help states realize efficiencies* (GAO 03-389). (Downloaded February 10, 2005, from http://www.gao.gov/cgi-bin/getrpt?GAO-03-389.) (Available from ERIC Document Reproduction Service, ERIC document no. ED477361.)

Gipps, C. (1999) Socio-cultural aspects of assessment. *Review of Research in Education*, *24*, 355-392.

Gitomer, D. H., & Duschl, R. (1995). Moving toward a portfolio culture in science education. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 299-325). Mahwah, NJ: Erlbaum.

Glaser, R. (1960). Christmas past, present, and future: A review and preview. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning: A source book* (pp. 23-31).Washington, DC: National Education Association.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519-521.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. M. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 625-670). Washington, DC: American Council on Education.

Goldberg, G. L., & Rosewell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance based instruction and classroom practice. *Educational Assessment, 6*, 257-290.

Gong, B., & Reidy, E. (1996). Assessment and accountability in Kentucky's school reform. In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 215-233). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Gordon, E. W., & Bonilla-Bowman, C. (1996). Can performance-based assessments contribute to the achievement of educational equity? In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 32-51). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Gotkin, L. G., & McSweeney, J. F. (1967). Learning from teaching machines. In P. C. Lange (Ed.), *Programmed instruction* (Sixty-sixth yearbook of the National Society for the Study of Education, Part II, pp. 255-283). Chicago: The National Society for the Study of Education, distributed by the University of Chicago Press.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan*, *80*, 662-666.

Haertel, E. H., & Calfee, R. C. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement*, *20*, 119-132.

Haertel, E. H., & Lorié, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 61-103.

Herman, J. L. (1997). Large-scale assessment in support of school reform: Lessons learned in the search for alternative measures. *International Journal of Educational Research, 27*, 395-413.

Herman, J. L. (2004). The effects of testing on instruction. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability* (pp. 141-166). New York: Teachers College Press.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J. L., & Golan, S. (1993). Effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, *12*(4), 20-25; 41-42.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527-535.

Kellaghan, T., & Madaus, G. (1991). National testing: Lessons for America from Europe. *Educational Leadership*, *49*(3), 87-93.

Lagemann, E. C. (1997). Contested terrain: A history of education in the United States, 1890-1990. *Educational Researcher*, *26*(9), 5-17.

Lange, P. C. (Ed.) (1967). *Programmed instruction* (Sixty-sixth yearbook of the National Society for the Study of Education, Part II). Chicago: University of Chicago Press.

LeMahieu, P. G., & Eresh, J. T. (1996). Coherence, comprehensiveness, and capacity in assessment systems: The Pittsburgh experience. In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 125-142). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Lindvall, C. M., & Cox, R. C. (1969). The role of evaluation in programs for individualized instruction. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means* (Sixty-eighth yearbook of the National Society for the Study of Education, Part II, pp. 156-188). Chicago: University of Chicago Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 84-103). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Madaus, G. F., Stufflebeam, D., & Scriven, M. S. (1983). Program evaluation: An historical overview. In G. F. Madaus, M. S. Scriven, & D. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 3-22). Norwell, MA: Kluwer Academic Publishers.

Mager, R. F. (1962). *Preparing instructional objectives*. Palo Alto, CA: Fearon Publishers.

Mager, R. F. (1975). *Preparing instructional objectives* (2nd ed.). Belmont, CA: Fearon Publishers.

Mager, R. F. (1984). *Preparing instructional objectives* (rev. 2nd ed.). Belmont, CA: Pitman Management and Training.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

National Commission on Excellence in Education [NCEE]. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Education Goals Panel. (1991). *The National Education Goals Report: Building a nation of learners*. Washington, DC: National Education Goals Panel. (For sale by the U.S. Government Printing Office, Superintendent of Documents.)

National Research Council (1983). *Education for tomorrow's jobs*. Committee on Vocational Education and Economic Development in Depressed Areas (S. W. Sherman, Ed.). Washington, DC: National Academy Press.

National Research Council (2000). *How people learn: Brain, mind, experience, and school* (expanded edition). Committee on Developments in the Science of Learning (J. D. Bransford, A. L. Brown, & R. R. Cocking, Eds.). Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment (J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC. National Academy Press.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state mandated testing programs on teaching and learning*. Boston: National Board on Educational Testing and Public Policy.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, R. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan*, *66*, 628-634.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, *6*, 1-9.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3-14.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Tech. Rep. No. 566). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Resnick, D. (1980). Minimum competency testing historically considered. *Review of Research in Education*, *8*, 3-29.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119-144.

Shavelson, R. J., Baxter, G. P., & Gao, X. H. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215-232.

Shepard, L. A. (1991a). Psychometricians' beliefs about learning. *Educational Researcher*, *20*, 2-16.

Shepard, L. A. (1991b). *Will national tests improve student learning?* (Paper presented at the American Educational Research Association Public Interest Invitational Conference, Accountability as a State Reform Instrument: Impact on Teaching, Learning, and Minority Issues and Incentives for Improvement, Washington, DC, June 5, 1991. Available as CSE Tech. Rep. No. 342, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation). UCLA: CRESST.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Skinner, B. F. (1953). *Science and human behavior*. New York: MacMillan.

Skinner, B. F. (1960). Teaching machines. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning* (pp. 137-158). Washington, DC: National Education Association.

Skinner, R. A. (2005, January 6). State of the states. *Education Week*, *24*(17), 77-106.

Smith, E. R., Tyler, R. W., & the Evaluation Staff. (1942). *Appraising and recording student progress* (The Adventure in American Education Series, Vol. III). New York: Harper & Bros.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, *10*(4), 7-11.

Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classroom* (CSE Tech. Rep. No. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stecher, B. M., Barron, S. L., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-1997 RAND survey of Kentucky teachers of mathematics and writing* (CSE Tech. Rep. No. 482). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stiggins, R. J. (2002) Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Terman, L. M. (1919). *The intelligence of school children*. Boston: Houghton-Mifflin.

Thorndike, E. L. (1910). The contribution of psychology to education. *Journal of Educational Psychology*, *1*, 5-12.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *The measurement of educational products*. (Seventeenth yearbook of the National Society for the Study of Education, Part II, pp. 16-24). Bloomington, Ill: Public School Publishing Company.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Varenne, H., & McDermott, R. (1999). *Successful failure: The school America builds*. Boulder, CO: Westview Press.

Walker, D. F., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research*, *44*, 83-111.

Washburne, C. W. (1925). A program of individualization. In C. W. Washburne (Chairman), *Adapting the schools to individual differences* (Twenty-fourth yearbook of the National Society for the Study of Education, Part II, pp. 257-272). Bloomington, IL: Public School Publishing Co.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* Madison, WI: National Institute for Science Education and Council of Chief State School Officers.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, *49*(8), 26-33.

Wiliam, D. (2004), *Assessing instructional and assessment practice: What makes a lesson formative?* Presentation at 2004 annual conference of the National Center for

Research on Evaluation, Standards and Student Testing (CRESST). Los Angeles, CA.

Wolf, D. P., & Reardon, S. F. (1996). Access to excellence through new forms of student assessment. In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 1-31). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).