Assessments and Accountability

CSE Technical Report 490

Robert L. Linn
CRESST/University of Colorado at Boulder

November 1998

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

Project 3.5 Coherence and Collaboration: The Equity and Technical & Functional Quality
Robert L. Linn, Project Director, University of Colorado at Boulder/CRESST

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

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

ASSESSMENTS AND ACCOUNTABILITY[1]

Robert L. Linn

Center for Research on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

Abstract

Uses of tests and assessments as key elements in five waves of educational reform during the past 50 years are reviewed. These waves include the role of tests in tracking and selection emphasized in the 1950s, the use of tests for program accountability in the 1960s, minimum competency testing programs of the 1970s, school and district accountability of the 1980s, and the standards-based accountability systems of the 1990s. Questions regarding the impact, validity, and generalizability of reported gains and the credibility of results in high-stakes accountability uses are discussed. Emphasis is given to three issues of currently popular accountability systems. These are (a) the role of content standards, (b) the dual goals of high performance standards and common standards for all students, and (c) the validity of accountability models. Some suggestions for dealing with the most severe limitations of accountability are provided.

Reform seems to be a constant part of the educational landscape. The details change frequently. Even the guiding philosophies and major themes may change from one reform to the next. At times a new reform involves a major shift or pendulum swing as one ideological camp gains ascendance over another. Sometimes a major shift may be supported by a legislative mandate or by policies adopted by state or local boards of education. The California mandates for whole-language instruction and, more recently, for phonics instruction provide obvious examples. A more balanced approach that combines features of phonics and literature-based instruction along the lines articulated in the recently released National Research Council report on reading (Snow, Burns, & Griffin,

[1] Based on a paper for the American Educational Research Association Career Research Contributions Award presented at the annual meeting of AERA, San Diego, April 16, 1998.

1998) would also involve a reform, albeit one that may be harder to sell as the simple quick fix that characterizes the rhetoric of many reform efforts.

Assessment and accountability have had prominent roles in many of the reform efforts during the last forty years. Testing and assessment have been both the focus of controversy and the darling of policy makers (Madaus, 1985).

There are several reasons for the great appeal of assessment to policy makers as an agent of reform.

First, tests and assessments are relatively inexpensive. Compared to changes that involve increases in instructional time, reduced class size, attracting more able people to teaching, hiring teacher aides, or programmatic changes involving substantial professional development for teachers, assessment is cheap.

Second, testing and assessment can be externally mandated. It is far easier to mandate testing and assessment requirements at the state or district level than it is to take actions that involve actual change in what happens inside the classroom.

Third, testing and assessment changes can be rapidly implemented. Importantly, new test or assessment requirements can be implemented within the term of office of elected officials.

Fourth, results are visible. Test results can be reported to the press. Poor results in the beginning are desirable for policy makers who want to show they have had an effect. Based on past experience, policy makers can reasonably expect increases in scores in the first few years of a program (see, for example, Linn, Graue, & Sanders, 1990) with or without real improvement in the broader achievement constructs that tests and assessments are intended to measure. The resulting overly rosy picture that is painted by short-term gains observed in most new testing programs gives the impression of improvement right on schedule for the next election.

Of course, tests and assessments come in many different forms and may be used in a variety of ways in accountability systems intended to improve education. It is important to identify features of assessment and accountability systems that influence both the trustworthiness of the information provided and the likely impact of the systems on educational practices and student learning. The purpose of this paper is to review some of those factors and to suggest a few

principles that may improve the trustworthiness of the information and enhance the positive effects of assessment and accountability while reducing some of the unintended negative side effects.

A review of several waves of reform since World War II that were based, in part, on assessment and accountability is provided in the next section. Salient features of assessment and accountability systems in current reform efforts are then discussed. Experience with the uses made in the past leads to some principles that are incorporated to varying degrees in current assessment and accountability systems. Other principles are suggested by the analysis of the systems that are currently in use or under development. Both sets of principles are summarized in a list of suggestions for assessment and accountability systems in the concluding section.

Five Decades of Assessment-Based Reform

Since World War II there have been several waves of reform involving test use. Both the roles that tests play in reform efforts and sometimes the nature of the tests have changed in each new wave of reform.

The influential writings of James B. Conant in the 1950s (e.g., 1953) provided a rationale for "universal elementary education, comprehensive secondary education, and highly selective meritocratic higher education" (Cremin, 1989, p. 22). Tests were seen as important tools to support the implementation of Conant's conceptualization of the educational system, both for purposes of selecting students for higher education and for identifying students for gifted programs within comprehensive high schools. In Conant's view, comprehensive high schools should provide students with a common core, but should also have differentiated programs. Indeed, as Cremin (1989) has noted, Conant believed that the preservation of "quality of education for the academically talented in comprehensive high schools" was a central problem for American education.

This vision of differentiated tracks within comprehensive high schools differs sharply from the current rhetoric of common high standards for all students. The current reality, however, is more congruent with the notion of differentiated tracks than it is with the current rhetoric. The difference between our rhetoric and the reality of differentiation in instructional offerings was

highlighted a decade ago in *The Underachieving Curriculum* (McKnight et al., 1987), where results from the Second International Mathematics Study (SIMS) were used to assess U.S. school mathematics curricula and performance. *The Underachieving Curriculum* documented that four quite different types of mathematics classes—called remedial, typical, enriched, and algebra classes—were prevalent in the U.S. What was taught and what was learned varied dramatically as a function of class type.

Corresponding to the great variation in class type was the finding in SIMS that the class component of variance accounted for almost half of the total variability in performance in the U.S., whereas the class component accounted for a much smaller fraction of the total variability in most other countries.

Recent reports based on the Third International Mathematics and Science Study (TIMSS) (Schmidt & McKnight, 1998) show that tracking in mathematics is still the norm by the eighth grade. The TIMSS analyses showed that almost 75% of the eighth-grade students were in schools that offered two or more distinct types of mathematics classes (Schmidt & McKnight, 1998). The corresponding percentages for Canada, England, and Germany ranged from about 10% to 20% whereas those for France, Korea, and Japan were between zero and 1%.

Contrary to some common assumptions, the between-school variance component for the U.S., although slightly larger than that for Japan and Korea, is actually somewhat smaller than it is in a number of countries (e.g., Canada, England, France), substantially so in the case of Germany. The class component was not estimated for the other countries, but because most schools in the other countries have only one type of mathematics class at age 13, one would expect the class component to be relatively small in the comparison countries. In the U.S., however, tracking results in a between-class component within schools that is larger than the between-school component.

A question that is unanswered is how much the variance would be reduced in the U.S. if tracking was eliminated at this age level in mathematics.

Elementary and Secondary Education Act

The tracking evident in Grade 8 is, of course, a reflection of differences in earlier educational experiences and achievement of students. Indeed, differences in achievement are evident when children begin first grade and increase with grade level. In recognition of the large disparities in educational opportunities

and in student performance, considerable attention was focused on compensatory education in the mid 1960s. The Elementary and Secondary Education Act (ESEA) of 1965 put in place the largest and most enduring of these federal efforts in this realm.

The congressional demands for evaluation and accountability for the funds distributed under Title I of ESEA as well as several other programs of that era proved to be a boon to test publishers. The testing demands of the Title I Evaluation and Reporting System (TIERS) (Tallmadge & Wood, 1981) contributed to a substantial expansion in the use of norm-referenced tests. Rather than administering tests once a year in selected grades, TIERS encouraged the administration of tests in both the fall and the spring for Title I students in order to evaluate the progress of students participating in the program. Although little use was made of the aggregate test results, these TIERS requirements relieved the pressure from demands for accountability for this major federal program.

In addition to increasing the use of standardized tests, TIERS led to a new reporting scale, the Normal Curve Equivalent (NCE). NCEs are simply normalized standard scores with a mean of 50 and a standard deviation of 21.06, which happens to be the standard deviation that makes NCEs coincide with National Percentile ranks at three points, namely 1, 50, and 99.

Nationally aggregated results for Title I students in Grades 2 through 6 showed radically different patterns of gain for programs that reported results on different testing cycles (Linn, Dunbar, Harnisch, & Hastings, 1982). Programs using an annual testing cycle (i.e., fall-to-fall or spring-to-spring) to measure student progress in achievement showed much smaller gains on average than programs that used a fall-to-spring testing cycle. The typical gain for the annual testing cycle reported by Linn et al. (1982), for example, was approximately 2 NCEs across Grades 2 through 6. The corresponding average gains for the fall-to-spring cycle, however, were between 6 and 9 NCEs for Grades 2 through 6.

Taken at face value, the preferred and more prevalent fall-to-spring testing cycle results painted quite a positive picture. Comparisons to the annual testing cycle results as well as comparisons to results of a number of large-scale evaluations of Title I, however, provided a rather compelling case for concluding that the fall-to-spring results were providing inflated notions of the aggregate effects of Title I programs. The best estimates from the era of the 1970s and early

1980s are that typical gains were on the order of magnitude found for the annual testing cycle, that is, something closer to 1 or 2 NCEs.

Linn et al. (1982) reviewed a number factors that together tended to inflate the estimates of gain in the fall-to-spring testing cycle results. These included such considerations as student selection, scale conversion errors, administration conditions, administration dates compared to norming dates, practice effects, and teaching to the test. Corruption of indicators is a continuing problem where tests are used for accountability or other high-stakes purposes. As discussed below in several other contexts, this tendency for scores to be inflated and therefore to give a distorted impression of the effectiveness of an educational intervention is not unique to TIERS. Nor is only of historical interest.

Several cautions for current assessment and accountability systems are suggested from the TIERS experience. Two obvious, but too frequently ignored, cautions are these: (a) Variations in which students get included in an assessment can distort comparisons of results for cohorts of students (e.g., those tested in the fall vs. those tested in the spring; and (b) reliance on a single test for repeated testing can distort instruction and lead to inflated and nongeneralizable estimates of student gains in achievement.

Minimum Competency Testing

In the 1970s and early 1980s minimum competency testing (MCT) reforms swiftly spread from state to state. In a single decade (1973-1983) the number of states with some form of minimum competency testing requirement went from 2 to 34. As the name implies, the focus was on the lower end of the achievement distribution. Minimal basic skills, while not easy to define, were widely accepted as a reasonable requirement for high school graduation. The new requirements were of great importance for some students but had little relevance for most students. Gains in student achievement were observed but they occurred mostly at low end of distribution. Moreover, questions were raised by some about generalizability of the observed gains.

For several reasons Florida was the focus of a great deal of attention in the MCT movement. Florida introduced a statewide MCT graduation requirement with a fairly tight time line for implementation. Florida's early results were used as examples of the positive effects of the program for students whose achievement lagged farthest behind expectations. Florida's results were also used

by detractors who emphasized differential passing rates for African American, Hispanic, and White students and the impact of the program on student dropout rates (Jaeger, 1989).

Federal District Court decisions in the Debra P. vs. Turlington (1981) case established precedents for many other states that continue to guide high-stakes requirements for individual students. The debate about opportunity to learn and whether it is fair to introduce high stakes for students without evidence that students have been provided with an adequate opportunity to learn the material on the test continues to be important for high-stakes requirements for students that have either been recently introduced or are now on the drawing board.

Figure 1 displays the passing rates on the first attempt of the Florida high school competency test by year of administration and racial/ethnic group. The percentages used to draw Figure 1 were obtained from the Florida Department of Education web site (http://www.firn.edu/doe/sas/sasshome.htm). The differential passing rates on first attempt were both dramatic and disturbing on the first administration in 1977. Roughly three fourths of the White students passed on the first attempt compared to slightly less than one fourth of the African American students and about three fifths of the Hispanic students. As can also be seen, however, the percent passing on the first try increased fairly sharply for all three racial/ethnic groups in year 2 of the program. Smaller increases followed in year 3 (1979) for all groups. With the exception of 1984, when there was a sharp increase for all three groups, and a couple of other minor dips and bumps, the trend is relatively flat for White and Hispanic students from 1979 to 1997. The passing rate on the first try for African American students, however, increased each year from the low of 23% in 1977 to an all-time high of 70% in 1984. Since 1984, however, there has been gradual erosion of the first-try pass rate for African Americans.

The pattern of early gains followed by a leveling off is typical not only of minimum competency testing programs but of several other high-stakes uses of tests. An unambiguous evaluation of the benefits remains elusive, however, because it is difficult to determine whether the gains are specific to the tests or whether they can be validly generalized to broader constructs the tests are intended to measure. A caution that might be taken from the MCT experience and results such as those in Figure 1 is that gains in the first few years following the introduction of a new testing requirement are generally much larger than

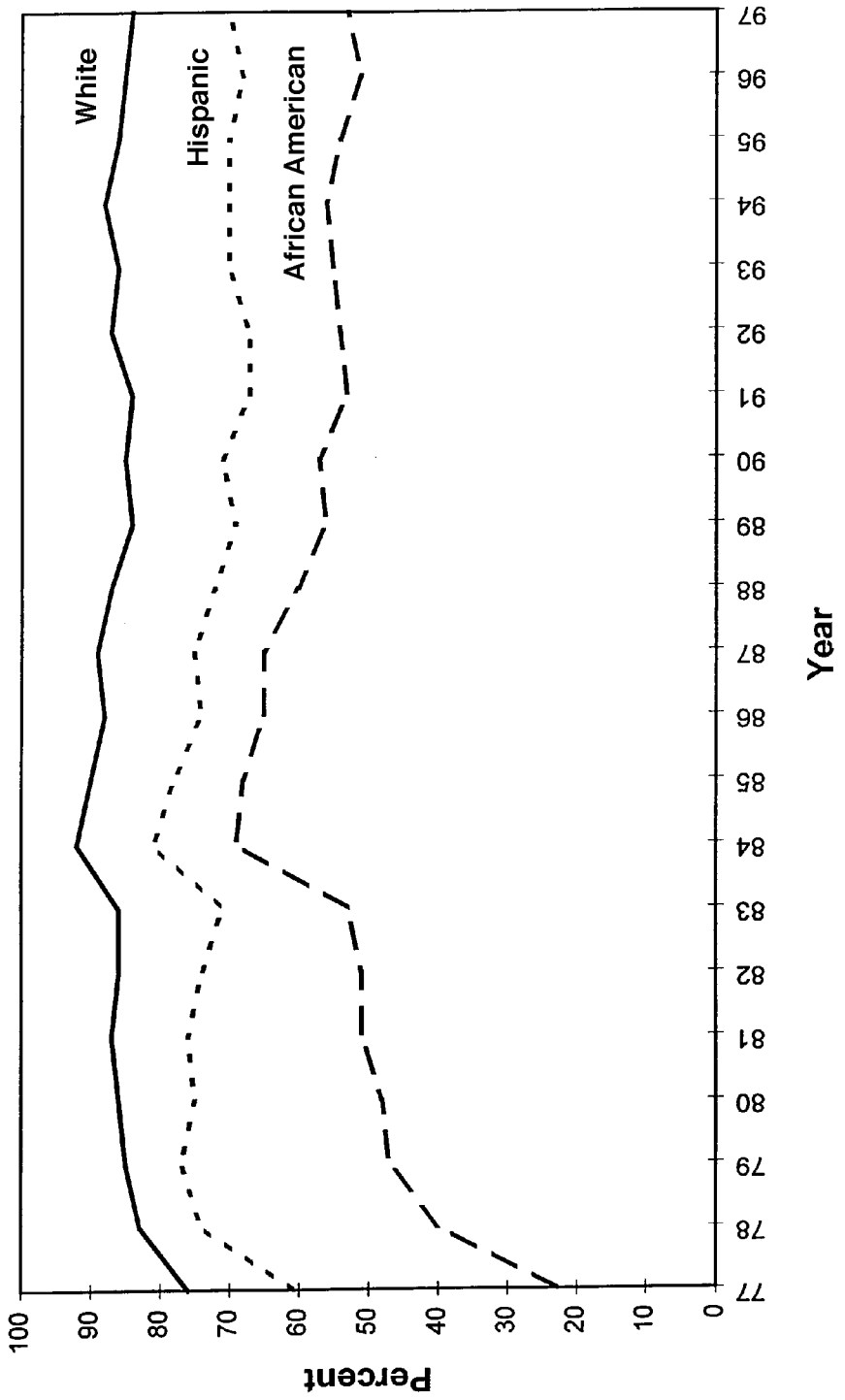

Figure 1. State percent passing on first try on Florida high school competency test by year and racial/ethnic group.

those achieved after the program has been in place for several years. This tendency raises questions about the realism of some accountability systems that put in place straight-line improvement targets over extended periods (e.g., twenty years).

Accountability Based on Standardized Tests

Overlapping with the minimum competency testing movement and continuing past the height of that movement into the late 1980s and early 1990s was an expansion of the use of test results for accountability purposes. Accountability programs took a variety of forms, but shared the common characteristic that they increased real or perceived stakes of results for teachers and educational administrators.

Although some states and districts contracted for or developed their own tests, the accountability systems of the 1980s relied heavily on published standardized tests. Upward trends in student achievement were reported by an overwhelming majority of states and districts during the first few years of accountability testing programs. A physician, John Cannell (1987), forcefully brought to public attention what came to be known as the Lake Wobegon effect (Koretz, 1988), that is, the incredible finding that essentially all states and most districts were reporting that their students were scoring above the national norm. Based on his review of the data, Cannell (1987) concluded that "standardized, nationally normed achievement test[s] give children, parents, school systems, legislatures, and the press inflated and misleading reports on achievement levels" (p. 3).

Most test publishers found increases in performance in their norms during the 1980s (Linn et al., 1990). If student performance were improving nationwide, then comparison of results to old norms would put a current average performance above the mean of the old norms. However, gains on NAEP were more modest than the gains found on most standardized tests, which raised doubts about the generalizability or robustness of the putative gains that were reported on standardized tests (Linn et al., 1990).

There are many reasons for the Lake Wobegon effect, most of which are less sinister than those emphasized by Cannell. Among the many reasons are the use of old norms, the repeated use of the same test form year after year, the exclusion of students from participation in accountability testing programs at a higher rate

than they are excluded from norming studies, and the narrow focusing of instruction on the skills and question types used on the test (see, for example, Koretz, 1988; Linn et al., 1990; Shepard, 1990). In each of the categories, practices range from quite acceptable to quite unacceptable. For example, the focusing of instruction on the general concepts and skills included in the test may be in keeping with the belief that the test corresponds to instructionally important objectives and considered acceptable, even desirable, practice. On the other hand, the narrow teaching of the specific content sampled by the test, or coaching in specific responses to test items would be widely condemned as unacceptable practice.

Whatever the reason for the Lake Wobegon effect, it is clear that the standardized test results that were widely reported as part of accountability systems in the 1980s were giving an inflated impression of student achievement. Striking evidence of this comes from trend results for states and districts that include a shift from an old to a new test. The pattern shown in Figure 2 is similar to ones observed repeatedly where a new test replaced one that had been in use by a state or district for several years. The sawtooth appearance in Figure 2 demonstrates the lack of generalizability of the apparent gains on a test that is reused for several years. Both common sense and a great deal of hard evidence indicate that focused teaching to the test encouraged by accountability uses of results produces inflated notions of achievement when results are judged by comparison to national norms.

Results reported by Koretz, Linn, Dunbar, and Shepard (1991) provide further evidence of lack of generalizability of accountability test results. Figure 3, which is adapted from results reported by Koretz et al. (1991), displays median Grade 3 test results in mathematics for a school district participating in that study. In year 1, when the district used standardized test 1, the median grade equivalent score for the district was 4.3. A new test, standardized test 2, was administered for accountability purposes for the first time in year 2 and used in each of the following years (3, 4, and 5). The sawtooth pattern similar to that in Figure 2 is clearly evident. That is, there is a sharp drop in scores the first year a new test is administered followed by gains on administrations in subsequent years.

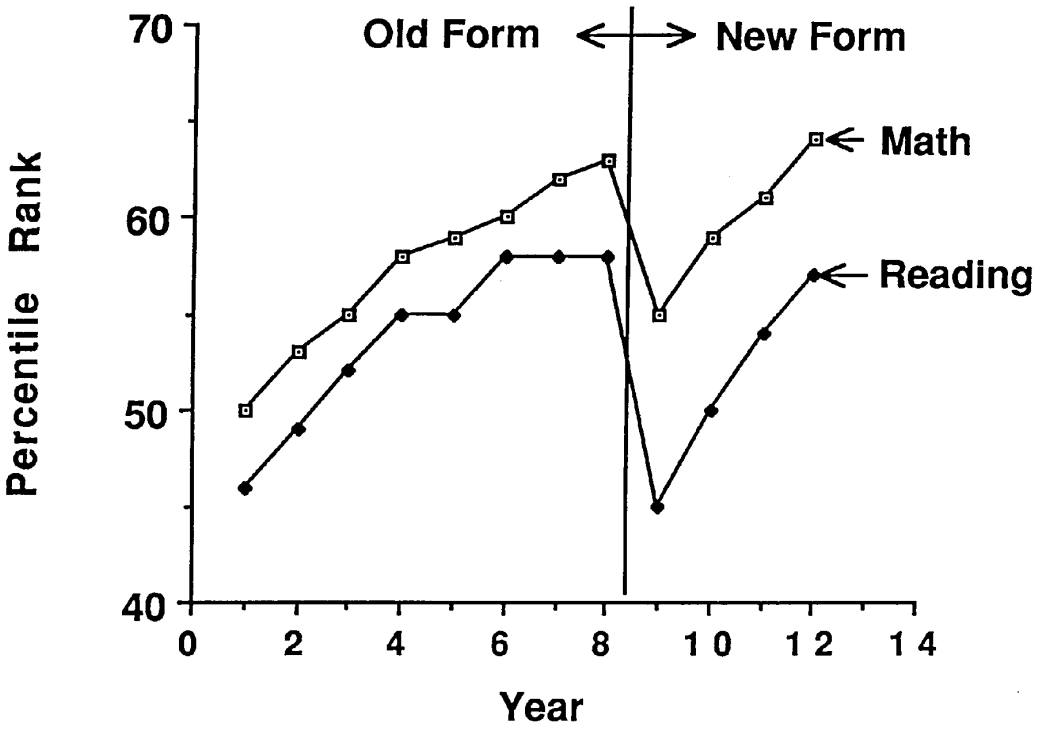

Figure 2. Trends in percentile rank of state means (based on Linn, Graue, & Sanders, 1990).

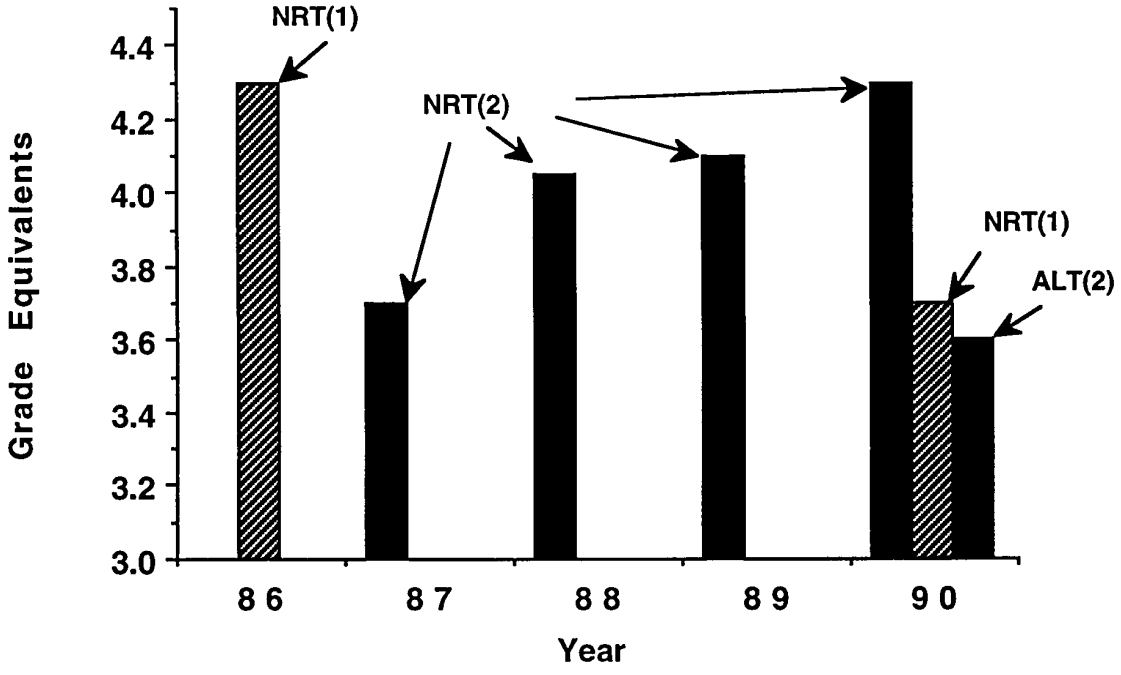

Figure 3. Inflated test scores (based on Koretz, Linn, Dunbar, & Shepard, 1991).

Koretz et al. (1991) administered standardized test 1 to a sample of students in the district in year 5. They also administered an alternative test that was constructed for the study to cover content defined by the district curriculum and the content of standardized test 2. Data collected in other districts were used to equate the alternate test to standardized test 2. As can be seen in Figure 3, the results for both standardized test 1 (formerly the district's operational test) and the alternate test are more in line with those for the first year's administration of standardized test 2 than with the concurrent administration of that test in year 5.

Results such as those shown in Figures 2 and 3 were used to make the case that standardized test results in high-stakes accountability systems were yielding inflated impressions of student achievement. Strong arguments were also advanced that high-stakes accountability uses of standardized tests had undesirable effects on teaching and learning because they led to a narrowing of the curriculum and an overemphasis on basic skills (e.g., Resnick & Resnick, 1992). One response has been to call for changes in the nature of assessments and the degree to which they are aligned with the types of learning envisioned in emerging content standards.

Salient Characteristics of Current Reform Efforts

The most recent wave of reform continues to emphasize accountability, but adds some significant new features. Although a number of other important features might be considered (e.g., the emphasis on performance-based approaches to assessment, the concept of tests worth teaching to, and the politically controversial and technically challenging issue of opportunity to learn), the focus below is on just three features. These are (a) the emphasis on the development and use of ambitious content standards as the basis of assessment and accountability, (b) the dual emphasis on setting demanding performance standards and on the inclusion of all students, and (c) the attachment of high-stakes accountability mechanisms for schools, teachers, and sometimes students.

Content Standards

A key feature of current reform efforts is the creation of standards. Standards are central to the Clinton administration's education initiative explicated in the *Goals 2000: Educate America Act. Goals 2000* is reinforced by the requirements for Title I evaluation stipulated in the *Improving America's Schools Act of 1994*. The federal government has encouraged states to develop

content and performance standards that are demanding. Standards are also a component of the President's proposal for a Voluntary National Test. Of course, standards-based reform reporting is also a central part of many of the state reform efforts, including states that have been using standards-based assessments for several years, such as Kentucky and Maryland, and states that have more recently introduced standards-based assessment systems, such as Colorado and Missouri. Indeed, states have been the key actors in standards-based reforms for a number of years. Given the current congressional press to reduce the federal role and give more flexibility and responsibility to states, it is reasonable to expect that the states will continue to be the key actors in standard-based reform efforts in the foreseeable future.

There is a great deal to be said about content standards. Indeed, a great deal has already been written about the strengths and weakness of content standards. State content standards have been reviewed and graded (e.g., *Education Week*, 1997; Lerner, 1998; Olson, 1998; Raimi & Braden, 1998). A review of the different perspectives brought to bear on state standards and the grades assigned to those standards is beyond the scope of this manuscript. But, it is worth acknowledging that content standards vary a good deal in specificity and in emphasis. The key point for present purposes, however, is that content standards can, and should, if they are to be more than window dressing, influence both the choice of constructs to be measured and the ways in which they are eventually measured. Moreover, it is critical to recognize first that the choice of constructs matters, and so does the way in which measures are developed and linked to the constructs. Although these two points may be considered obvious, they are too often ignored.

Table 1 provides one simple example of the fact that choice of constructs to be measured and used to hold students or teachers accountable matters. The table simply reports the percentage of male and female students who score at the proficient level or higher on four recent NAEP assessments. As can be seen, the gender difference in percentage passing based on the NAEP proficient level criterion varies considerably by subject. Sizable discrepancies could be displayed with other cut scores or with other subpopulations of students being contrasted. The point, which was made in much greater detail and with a wealth of supporting evidence by Willingham and Cole (1997), however, is that construct choice matters.

Table 1

Choice of Constructs Matters. The Percentage of Students At or Above the NAGB Proficient Level on NAEP at Grade 12 by Gender and Subject

Subject	Males	Females	Difference (M-F)
Geography (1994)	32	22	10
History (1994)	12	9	3
Mathematics (1996)	18	14	4
Reading (1994)	29	43	-14

Which content areas are assessed also makes a difference in accountability systems. It is difficult to compare the performance in different subjects. The fact that more students pass an English language arts test than pass a mathematics test, for example, could just as easily be due to differences in the rigor of the assessment or differences in where the standards are set in the different content areas as to fundamental differences in achievement. Trend data, however, can be used to show differences between content domains in gains or losses in performance. Even at the level of global subject areas such as mathematics, reading, and science some differences can be seen in long-term trends.

The results shown in Figure 4 are based on means reported in Tables A.17, B.17, and C.17 in the *NAEP 1996 Trends in Academic Progress* (Campbell, Voelkl, & Donahue, 1997). As is shown in Figure 4, the NAEP long-term trends for 13-year-olds in mathematics, reading, and science have had somewhat different trajectories. National means used to prepare Figure 4 were obtained from tables A.17, B.17, and C.17 of the appendices in Campbell et al. (1997). None of the trend lines is particularly steep, but the increases of roughly a quarter of a standard deviation in science for the 15 years between 1977 and 1992 and of almost a third of a standard deviation in mathematics over the 18 years between 1978 and 1996 are nontrivial. In contrast, the long-term trend for reading is quite flat with an increase of only about one-tenth of a standard deviation in the 25 years between 1971 and 1996. Moreover, as others have noted, the trends are not the same across subpopulations defined by gender, race/ethnicity, or type of community (e.g., Campbell et al., 1997; Jones, 1984).

Trends for subscales within broad content areas also reveal potentially important differences. This is evident in Figure 5, which shows a marked

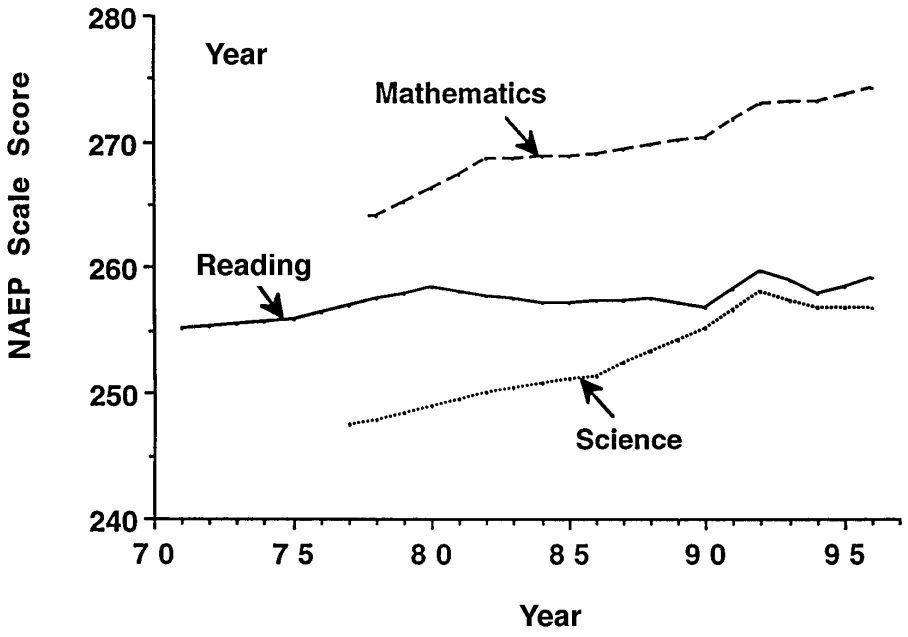

Figure 4. NAEP long-term trend lines for age 13 students in three subjects (based on Campbell, Voelkl, & Donahue, 1997).

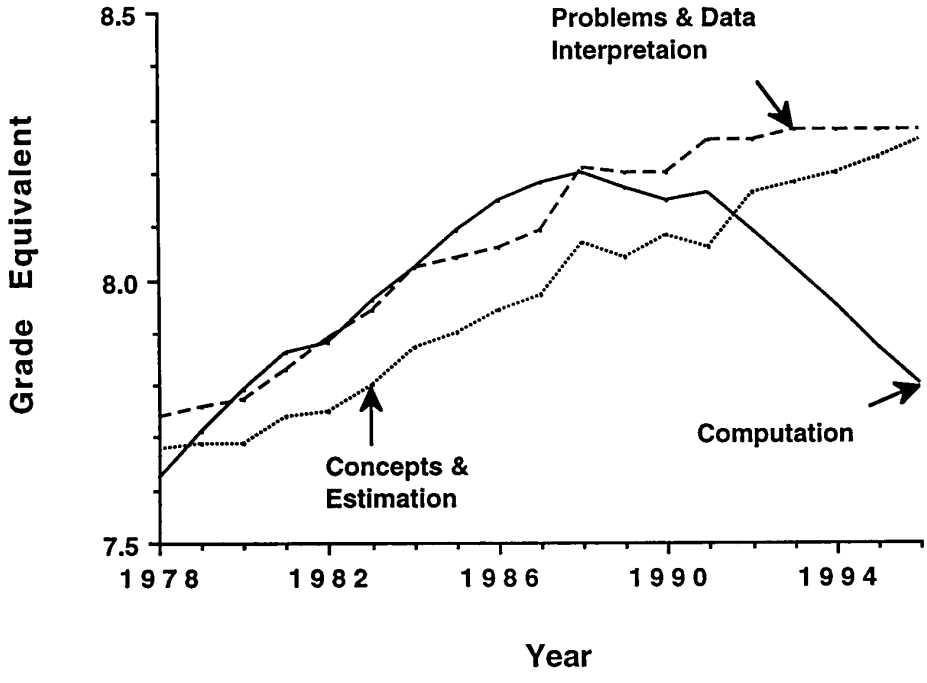

Figure 5. Divergent trends in Grade 8 ITBS mathematics subscores for the state of Iowa (H.D. Hoover, personal communication, August 1997).

difference between the recent trend for the computation subtest and the trends for the concepts and estimation and the problems and data interpretation subtests of the Iowa Tests of Basic Skills (ITBS) for the state of Iowa. Test subscale means used to produce Figure 5 were provided by H. D. Hoover (personal communication, August 1997).

Although results such as those in Figure 5 might be used as ammunition by critics of the NCTM *Standards* (1989) who argue that basic computational skills are given short shrift by the *Standards*, they are presented here to make two different points. First, we need to track more than global scores such as "mathematics" with the nation's primary achievement monitor, NAEP. We should have better information than we currently have at the national level on the components of mathematics and other subjects to monitor trends such as those shown on the ITBS mathematics subtests in Iowa. Note that the general trend in Figure 5 between 1978 and 1996 on the ITBS for two of the subtests is quite congruent with the gradual upward trend shown in Figure 4 for NAEP mathematics during the same time period. Possibly as important as the information about the trend for mathematics as a whole, however, may be the contrast of the trends for the subtests and what is made of the apparent differences.

Second, the ITBS differences in trends for subtests provide useful evidence that the ways in which content standards are used to determine the constructs to be measured and the specific assessment tasks to be used are critical. Whether computation is weighted heavily or lightly in the overall assessment is important. This is so despite the fact that using the usual individual differences approach to evaluating the distinctions among dimensions would undoubtedly yield quite high correlations among the subtests. High correlations at a given point in time do not tell the whole story.

For decades, test publishers have argued that subscores can be useful in suggesting areas or relative strengths and weaknesses for students. Although that claim is subject to debate, it seldom even gets considered when aggregate results are used either to monitor progress (e.g., NAEP) or for purposes of school, district, or state accountability. Differences for aggregates in relative performance on content strands may be more revealing, however, than differences on global scores. Multiple measures are needed for monitoring and accountability systems.

Performance Standards and All Students

In addition to content standards, the current reform efforts place great stock in the dual goals of high performance standards and the inclusion of all students. Performance standards, while independent of content standards, add additional considerations. In particular, performance standards are supposed to specify "how good is good enough."

There are at least four critical characteristics of performance standards. First, they are intended to be absolute rather than normative. Second, they are expected to be set at high, "world-class" levels. Third, a relatively small number of levels (e.g., advanced, proficient) are typically identified. Finally, they are expected to apply to *all*, or essentially all, students rather than a selected subset such as college-bound students seeking advanced placement.

A reasonable question that generally goes unanswered is whether the intent is to aspire not just to high standards for all students, but to the *same* high standards for *all* students. And, moreover, to do so on the same time schedule (e.g., meet reading standards in English at the end of Grade 4) for all students. It is quite possible to have high standards without the standards being common for all students. High standards of performance for a given grade level do not necessarily mean common standards for all students. For example, setting high standards for a student based upon the student's IEP may not lead to an expectation of a proficient score on the fourth-grade reading test—whether the state's own fourth-grade reading test, a standardized test, or the proposed Voluntary National Test that is used to define the proficient standard. Similarly, an English language learner may more appropriately be tested in reading in Spanish at the fourth grade with expectations of achieving proficient performance in reading in English at a later grade.

The recent strategy in Washington, DC, seems to have been to shame states into getting their performance standards in line with the national standards—by which the federal proponents of this position mean the standards established on NAEP by the National Assessment Governing Board (NAGB). Figure 6 is an example of results that Secretary Riley has used in trying to make the case that state standards are not as challenging as the proficient-level standards established by NAGB. Figure 6 is adapted from a figure accompanying Secretary Riley's statement before the House Subcommittee on Early Childhood, Youth and

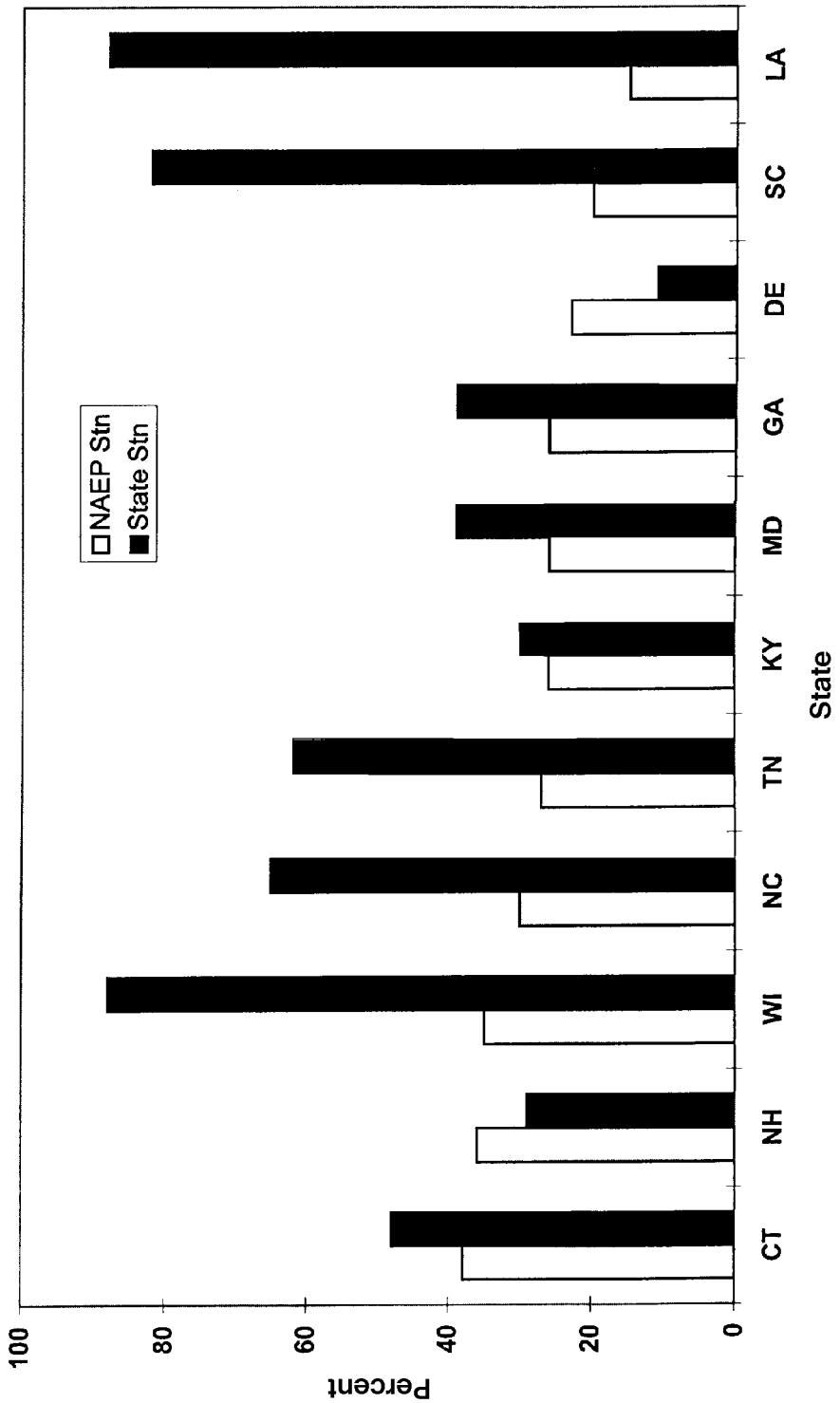

Figure 6. NAEP percent Proficient or Above vs. percent meeting state standards on states' own assessments (based on Secretary Riley's Statement to Congress (http://www.ed.gov/Speeches/04-1997/970429.html).

Families Committee on Education and the Workforce, Tuesday, April 29, 1997, which is available on the Web (http://www.ed.gov/Speeches/04-1997/970429.html). It is based on work reported by the Southern Regional Education Board, which can be found on the Web at (http://www.sreb.org/MiscDocs/set_stand.html) and which lists as sources the U.S. Department of Education, state departments of education, and the National Education Goals Panel.

Note the discrepancy between the results for the state and NAEP standards for Wisconsin. The 35% of Wisconsin students who were at or above the proficient standard for fourth-grade reading on NAEP compares reasonably well to other states or to the national figure of 36% proficient in 1994. The 35% meeting the NAEP proficient standard, however, is well out of line with the 88% of Wisconsin students that are reported to meet the state standard on their third-grade reading test. Although the two sets of results appear grossly out of line on the surface, they are only what is to be expected if the difference in purpose of the two sets of standards is considered.

According to NAGB, the proficient level on NAEP is intended to represent "solid academic performance." It is supposed to indicate that "students reaching this level have demonstrated competency over challenging subject matter" (from the description of achievement levels on NAGB web site, http//www.nagb.org/). Some, myself included, would say that the proficient standard is an ambitious standard intended to encourage greater effort. The Wisconsin state standard was set with a different purpose in mind. As described on the Wisconsin Department of Public Instruction Web site (http://www.dpi.state.wi.us/dpi/spr/3wrct97.html), "The Wisconsin Reading Comprehension Test . . . allows districts to evaluate their primary school reading programs in comparison to a statewide performance standard. It also helps identify marginal readers who may need remediation." A standard signifying "solid academic performance" is quite different from one intended to provide help in identifying "marginal readers who may need remediation." Implying that that the two standards should yield similar percentages is misleading and may undermine the credibility of performance standards.

Moving from Grade 4 reading to Grade 8 mathematics assessment and from state comparisons to international comparisons, a case can be made, as has been widely publicized, that students in some other countries have much higher

average performance than students in the U.S. It is also worth noting, however, that there is substantial variability in performance in all countries.

Figure 7 displays box-and-whisker plots for the U.S. and six other selected countries. Percentile points used to construct Figure 7 were obtained from Table E.1 of the Appendix E of Beaton et al.'s *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)* (1996). Although the distributions for Japan and Korea are substantially higher than the distribution for the U.S., there is a large spread in all countries. Using an arbitrary standard of the U.S. 75th percentile, substantially more than a quarter of the students in Japan and Korea would fail. The two horizontal lines are unofficial approximations to where the NAEP cut scores for the proficient and basic achievement levels fall on the TIMSS scale. To get those lines, it was simply assumed that the percent proficient or above or percent basic or above would be the same for the U.S. students on the eighth-grade reading test in TIMSS as it was on the 1996 eighth-grade mathematics on NAEP. Greater precision than that is not needed to make it clear that while a large percentage of students would meet either standard in some other countries, a substantial fraction of students would nonetheless fall far short of the standard in all six countries used in this comparison. This is obviously true for the proficient level, but is also true even for the basic level.

It is one thing to strive for high standards. It is another to enforce them as seems to be implied by a great deal of political rhetoric. President Clinton, like many politicians at the state and local levels, has called for an end to social promotion. In President Clinton's *Call to Action for American Education in the 21st Century*, for example, it is stated that "today, only a handful of states in the country require young people to demonstrate what they've learned in order to move from one level of school to the next. Every state should do this and put an end to social promotion" (quotation taken from the Web at http://www.ed.gov/updates/PresEDPlan/part2.html). The plea for the Voluntary National Test, for example, argues that it will help end social promotion. Two points are worth emphasizing in response to the plea. First, given the fact that a large proportion of students are retained in grade at least once during Grades 1 through 12 (Alexander, Entwisle, & Dauber, 1995; Shepard, Smith & Marion, 1996), it simply is *not* true that promotion is based only on years of education or social considerations (Huebert & Hauser, 1998). Second, any

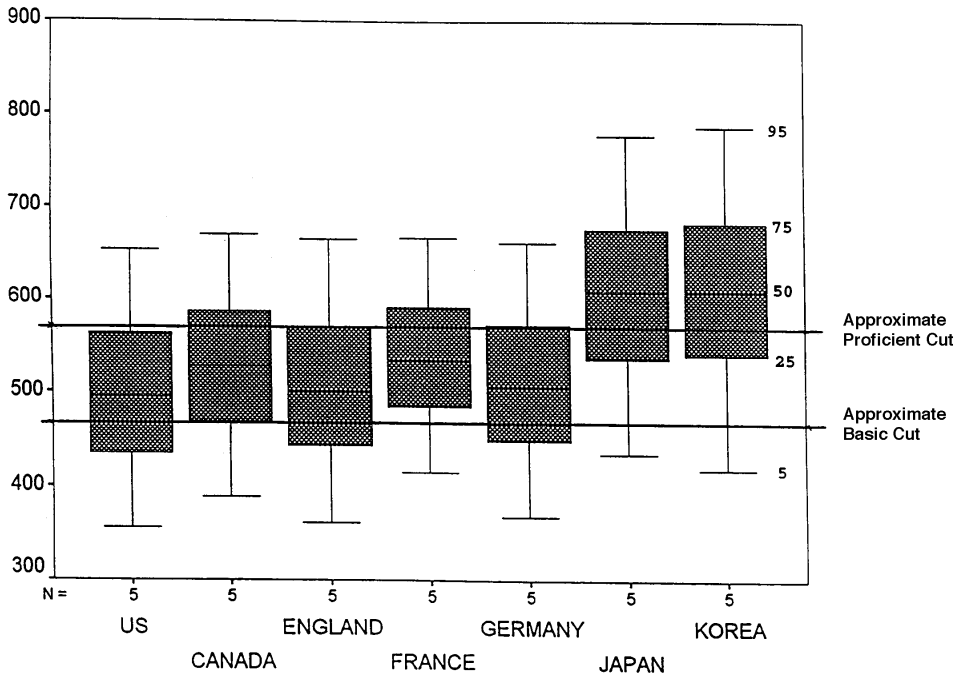

Figure 7. TIMSS Grade 8 mathematics results for selected countries (based on Beaton et al., 1997).

of the high standards that are now being set around the country and used for NAEP at the national level would fail an extremely large fraction of the students. As was just seen, that would be true even in countries such as Japan or Korea using Grade 8 mathematics cutoffs comparable to the NAEP proficient level or even the lower basic level. The story for some states and districts in the U.S. is even more sobering. On the 1996 NAEP Grade 8 mathematics assessment, for example, approximately 80% of the students in the District of Columbia scored in the Below Basic category, and only one student in 20 would meet the Proficient standard (see Table 3.2 in Reese, Miller, Mazzeo, & Dossey, 1997).

Coffman (1993) summed up the problems of holding common high standards for all students as follows. "Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents" (p. 8). Although this statement runs counter to the current zeitgeist, and may not even be considered politically correct, it seems to me a sensible conclusion that is consistent with both evidence and common sense. It should not be misinterpreted, however, to mean that one should not have high standards for

all students, but that is not the same as common standards for all, especially when they are tied to a lock step of age or grade level. Neither should agreement with Coffman's conclusion be misinterpreted as supporting placement of some students into tracks with watered-down, basic-skills instruction while others are provided with enriched experiences in keeping with the intent of ambitious content standards.

It is, as Glass (1978) said, "wishful thinking to base a grand scheme on a fundamental, unsolved problem" (p. 237). The problem of setting standards remains as much a fundamental, unsolved problem today as it was 20 years ago. Despite the emphasis on absolute judgments in terms of fixed standards that are demanded, desires for comparisons continue to surface. Sometimes these comparisons are made implicitly in descriptions of the standards as "world-class." Often, however, the comparisons are explicit in requirements that state performance be compared to national achievement, or to achievement of other states through NAEP or some other means, or to that of other nations through international assessments such as TIMSS). In California, districts are currently being encouraged to set as a target for Title I that 90% of their students score above the national median on a standardized test within the next 10 years. Such a notion of a common high standard for all students sounds a warning that although the Lake Wobegon effect may be largely forgotten, it is not necessarily gone.

High-Stakes Accountability

The third and final assessment-related characteristic of the current reform efforts to be discussed here concerns the attachment of high-stakes accountability mechanisms for schools, teachers, and sometimes students. The use of student performance on tests in accountability systems is not new. Examples of payment for results—such as the flurry of performance contracting in the 1960s—and other ways of using performance as the basis for holding school administrators, teachers, and students accountable can be found cropping up and fading away over many decades. What is somewhat different about the current emphasis on performance-based accountability is its pervasiveness.

"What is new is an increasing emphasis on student performance as the touchstone for state governance" (Elmore, Abelmann, & Fuhrman, 1996, p. 65). Student achievement is not only being used to single out schools that require

special assistance but, as in the case of Kentucky, to provide cash incentives for improvements in performance. Moreover, "the focus on performance has led to using outcome data, such as test scores and dropout rates, as criteria for accreditation" (Elmore et al., 1996, p. 66).

The intent of this emphasis on outcomes is clear. Elmore et al. (1996) described the underlying rationale as follows.

> In principle, focusing on students' performance should move states away from input regulations—judging schools based on the number of books in the library and the proportion of certified staff, for example—toward a model of steering by results—using rewards, sanctions, and assistance to move schools toward higher levels of performance. In other words, the educational accountability should focus schools' attention less on compliance with rules and more on increasing learning for students. (p. 65)

Despite the increased popularity of high-stakes accountability systems, there are several fundamental questions that have plagued efforts to build assessment-based school building accountability systems for decades (Dyer, 1970; Dyer, Linn, & Patton, 1969). The questions could be categorized in a variety of ways, but many of them fall conveniently into one of three categories. Those are questions related to (a) the student assessments, (b) the accountability model, and (c) the validity, impact, and credibility of the system.

Assessments. Much of the above discussion of assessments is relevant to their use in accountability systems. In particular, the point that the choice of constructs matters needs to be emphasized. Content areas assessed for a high-stakes accountability system receive emphasis while those that are left out languish. Within a content area, the emphasis given to different subareas matters. Consider in this regard the different trends for the different subscores of the ITBS in Iowa that was discussed earlier.

Meyer (1996) has argued that "in a high-stakes accountability system, teachers and administrators are likely to exploit all avenues to improve measured performance. For example, teachers may 'teach narrowly to the test.' For tests that are relatively immune to this type of corruption, teaching to the test could induce teachers and administrators to adopt new curriculums and teaching techniques much more rapidly than they otherwise would" (p. 140).

It is easy to agree, in principle, with Meyer's argument. It is unclear that there is either the know-how or the will, however, to develop assessments that

are sufficiently "immune to this type of corruption." It is expensive to introduce a new, albeit well-equated, form of a test on each new administration. Frequent new forms of tests are used for high-stakes programs such as college and graduate or professional school admissions testing where the individual test taker bears the cost. Still there are complaints about test coaching or narrow test preparation for such tests. Moreover, if ambitious performance-based tasks are added to the mix, still greater increases in costs will result. We should not expect, however, inexpensive tests designed for other low-stakes purposes to withstand the pressures now being placed on them by high-stakes accountability systems.

Accountability model. The current accountability model applications using hierarchical linear models in states such as Tennessee, North Carolina, and South Carolina and districts such as Dallas use highly sophisticated statistical machinery (see, for example, Burstein, 1980; Clotfelter & Ladd, 1996; Meyer, 1996; Raudenbush & Byrk, 1986; Sanders & Horn, 1994; Willms & Raudenbush, 1989). But, the sophisticated statistics do not resolve questions about what data the basic model should employ. Some possibilities include current status, comparisons of cross-sectional cohorts of students at different grades in the same year, comparisons of cross-sectional cohorts in a fixed grade from one year to the next, longitudinal comparisons of school aggregate scores without requiring matched individual data, and longitudinal comparisons based only on matched student records. When looking at changes over years, whether using aggregate scores for cross-sectional cohorts or matched or unmatched longitudinal data, there are questions about the model. Should simple change scores be used or some form of regression-based adjustment? And, if regression-based adjustments are used, what variables should be included as predictors? In particular, should measures of socio-economic status be used in the adjustments?

Variations on most of these themes can be found in models used around the country. Although most proponents of performance-based accountability systems with systematic rewards and sanctions would prefer another approach, current status scores remain the most commonly reported approach. Where change in performance is used, there are several variations in use. Comparisons of time-lagged cross-sectional cohorts, for example, are used as the basis for the Kentucky accountability system (see, for example, Gusky, 1994). Longitudinal results using hierarchical regression procedures are used in places such as Tennessee (e.g., Sanders & Horn, 1994) and Dallas.

There is considerable evidence that the choice of data source and choice of summary statistic matter a good deal. This is illustrated by the correlations reported in Table 2. These results were selected from a study by Clotfelter and Ladd (1996) in which they used nine different analytical models to analyze data from South Carolina schools (Richards & Sheu, 1992). The five approaches for which correlations are reported in Table 2 are (a) school means, that is, a current status measure without any adjustments for SES or past performance, (b) simple mean gain scores from one year to the next for matched longitudinal cases, (c) the mean gain obtained by subtracting the mean fourth-grade score for a school from the previous year from the mean fifth-grade score in the current year, that is, unmatched longitudinal, (d) the school gain index actually used by South Carolina, which is based upon student residual scores obtained from a multivariate regression of current year scores on scores from the prior year, and (e) residual scores based upon the model used in North Carolina, which includes adjustments for SES.

With the possible exception of the correlation between the matched and unmatched gain scores, the correlations in Table 2 are not high enough to suggest that the choice of models is a matter of indifference. Some of the correlations are far too low to conclude that the approaches could be used interchangeably. Particularly notable are the low correlations in the last row, which involve the relationship of the only SES adjusted index with the other indices.

Table 2

Correlations Among Alternate Indices of School Performance (Clotfelter & Ladd, 1996)

Index	(1) Mean score	(2) Gain (M)	(3) Gain (UM)	(4) SC SGI	(5) NC SES Adj
(1) Mean score	1.00				
(2) Gain (Match)	.42	1.00			
(3) Gain (UnMatch)	.36	.94	1.00		
(4) SC SGI	.51	.89	.86	1.00	
(5) NC SES Adj	.58	.47	.44	.22	1.00

Note. SC = South Carolina. SGI = School gain index. NC = North Carolina.

Knowing that SES and prior achievement adjustments make nontrivial differences leaves unanswered the question about which model is most appropriate. The issue was well framed by Elmore et al. (1996):

> One side of this issue . . . argues that schools can fairly be held accountable only for factors that they control, and therefore that performance accountability systems should control for or equalize student socioeconomic status before they dispense rewards and penalties. . . . The other side of the issue argues that controlling for student background or prior achievement institutionalizes low expectations for poor, minority, low-achieving students. (pp. 93-94)

Although somewhat less controversial than adjustments for SES variables, similar arguments can be made regarding adjustments for prior achievement since, in effect, such adjustments establish lower expectations for schools with students whose prior achievement is low than for schools where it is high. Kentucky's approach to this dilemma has been to set a common goal for all schools by the end of 20 years, thus establishing faster biennial growth targets for initially low-scoring schools than initially high-scoring schools (Gusky, 1994). This approach is an appealing compromise between the extremes represented by using status measures and using residuals after adjusting for both prior achievement and family background characteristics.

The conceptual appeal of value-added models must be weighed not only against the concerns articulated by Elmore et al. (1996) but also against practical concerns. To reduce problems of attrition due to student mobility and to produce timely results, Meyer (1996) strongly recommends "annual testing at each grade" (p. 141) and the collection of data on family and community characteristics for all students. This recommendation not only imposes a substantial testing and data collection burden, but is likely to lead to practices of test reuse that exacerbate problems of teaching to the test in the narrow and undesirable sense. It is also likely to lead to the use of inexpensive tests that are less ambitious than proponents of standard-based reform are seeking.

Validity, impact, and credibility. None of the arguments about measures and models that have been discussed directly addresses what is perhaps the most important question. Have the assessment-based accountability models that are now being used or being considered by states and districts been shown to improve education? Unfortunately, it is difficult to get a clear-cut answer to this simple question. Certainly, there is evidence that performance on the measures

used in accountability systems increases, but that was true in the days of the Lake Wobegon effect concerns with standardized tests. Comparative data are needed to evaluate the apparent gains. NAEP provides one source of such comparative data.

Figure 8 displays an example using trends for the Maryland Assessment in Mathematics at Grades 3 and 5. The Maryland results used to produce Figure 8 were obtained from the Maryland Department of Education Web site (http://www.msde.state.md.us/), and the NAEP results were obtained from Table 3.2 in the *NAEP 1996 Mathematics Report Card for the Nation and the States* (Reese et al., 1997). The two lines in the middle of the graph plot the trends in the percentage of third- and fifth-grade students who perform at the satisfactory level on the state mathematics assessment. Those lines are bracketed by the percentages of Maryland fourth-grade students who scored at or above the basic and proficient levels on the 1992 and 1996 NAEP assessments in mathematics. There is no reason to expect prefect correspondence of the

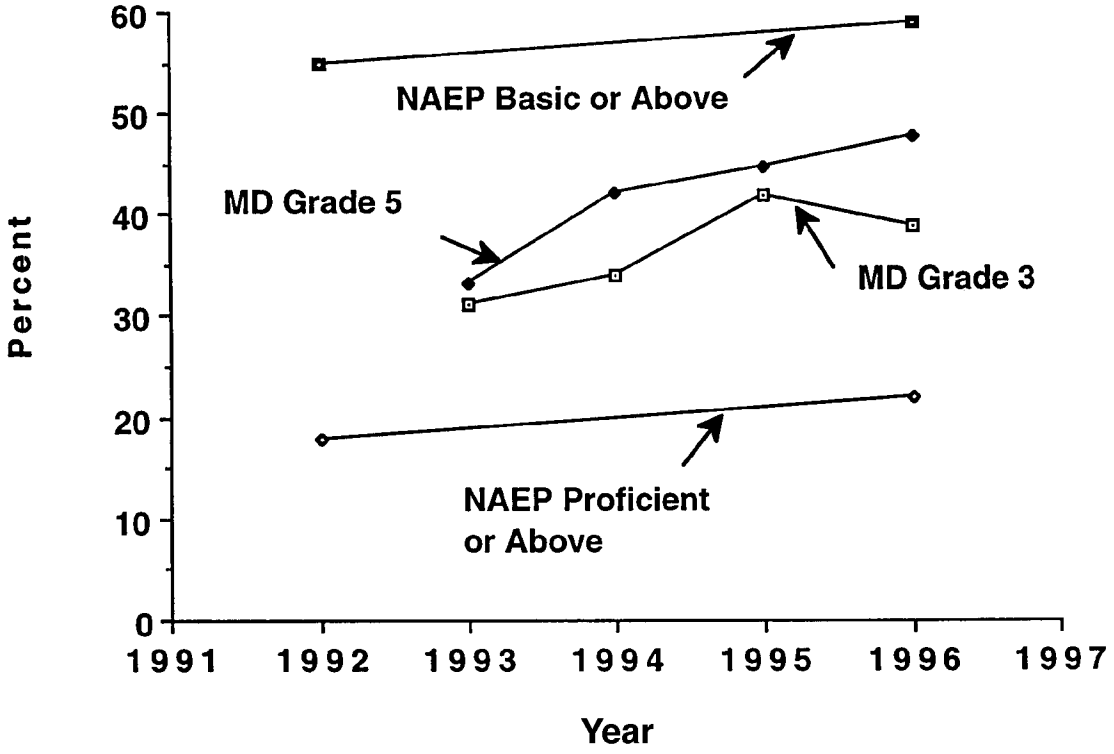

Figure 8. Percentages of Maryland students scoring Satisfactory on the Maryland assessment in mathematics at Grades 3 and 5 and the percentage of Maryland students at or above the Basic and Proficient levels on NAEP at Grade 4 by year.

trends. The gradual and reasonably similar upward trends shown by all four lines support the notion that these Maryland gains are reasonably generalizable to other measures and not simply providing an inflated impression of improvement.

Comparisons to state NAEP results similar to those illustrated in Figure 8 will sometimes provide evidence that generally confirms trends found on state assessment. In other cases, the trends for a state's own assessment and NAEP will suggest contradictory conclusions about the changes in student achievement. Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence about the validity of claims about student achievement.

Conclusion

As someone who has spent his entire career doing research, writing, and thinking about educational testing and assessment issues, I would like to conclude by summarizing a compelling case showing that the major uses of tests for student and school accountability during the past four decades have improved education and student learning in dramatic ways. Unfortunately, that is not my conclusion. Instead, I am led to conclude that in most cases the instruments and technology have not been up to the demands that have been placed on them by high-stakes accountability. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high-stakes accountability uses often outweigh the intended positive effects.

In a paper entitled "A King Over Egypt, Which Knew Not Joseph" Coffman (1993) used the story of Joseph and the Pharaohs as an analogy for policy makers who "know not" the advice of five educational Josephs. Although he argued that external tests are valuable for charting student progress and as an external comparison to the day-to-day observations of educators, Coffman concluded that "comparing averages among schools, systems, or states on any test is inherently unfair because it is not possible to separate school effects from effects resulting from nonschool factors" (1993, p. 8). Only Pollyanna could conclude that high-stakes accountability uses of assessments will wither away in the face of this statement. Nonetheless, it is worth arguing for more modest claims about uses

that can validly be made of our best assessments and warning against the overreliance on them that is so prevalent and popular.

It is toward this end that the following seven suggestions based on analyses discussed above are offered as ways of enhancing the validity, credibility, and positive impact of assessment and accountability systems while minimizing their negative effects.

1. Provide safeguards against selective exclusion of students from assessments. This would reduce distortions such as those found for Title I in the fall-spring testing cycle. One way of doing this is to include all students in accountability calculations.
2. Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years. Getting by on the cheap will likely lead to both distorted results (e.g., inflated, nongeneralizable gains) and distortions in education (e.g., the narrow teaching to the test).
3. Don't put all of the weight on a single test. Instead, seek multiple indicators. The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.
4. Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all.
5. Consider both value added and status in the system. Value added provides schools that start out far from the mark a reasonable chance to show improvement, while status guards against "institutionalizing low expectations" for those same students and schools.
6. Recognize, evaluate, and report the degree of uncertainty in the reported results.
7. Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

References

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1995). *On the success of failure: A reassessment of the effects of retention in the primary grades*. New York: Cambridge University Press.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. C. Berliner (Ed.), *Review of research in education* (pp. 158-223). Washington, DC: American Educational Research Association.

Campbell, J. R., Voelkl, K. E., & Donahue, P. L. (1997). *NAEP 1996 trends in academic progress*. Washington, DC: National Center for Education Statistics.

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

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 23-63). Washington, DC: The Brookings Institution.

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. *Educational Measurement: Issues and Practice*, *12*(2), 5-8.

Conant, J. B. (1953). *Education and liberty: The role of schools in a modern democracy*. Cambridge, MA: Harvard University Press.

Cremin, L. A. (1989). *Popular education and its discontents*. New York: Harper & Row.

Debra P. v. Turlington, 644 F.2d 397, 6775 (5th Cir. 1981).

Dyer, H. S. (1970). Toward objective criteria of professional accountability in schools of New York City. *Phi Delta Kappan*, *52*, 206-211.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal*, *6*, 591-605.

Education Week. (1997, January 22). Quality counts: A report card on the condition of public education in the 50 states. *A Supplement to Education Week, 16*.

Elmore, R. F., Abelmann, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 65-98). Washington, DC: The Brookings Institution.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237-261.

Gusky, T. R. (Ed.). (1994). *High stakes performance assessment: Perspectives on Kentucky's reform*. Thousand Oaks, CA: Corwin Press.

Huebert, J. P., & Hauser, R. M. (Eds.). (1998). *High-stakes testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: Macmillan.

Jones, L. V. (1984). White-black achievement differences: The narrowing gap. *American Psychologist, 39*, 1207-1213.

Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15, 46-52.

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171-195). Washington, DC: National Research Council.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lerner, L. S. (1998). *State science standards: An appraisal of science standards in 36 states*. Washington, DC: Thomas B. Fordham Foundation.

Linn, R. L., Dunbar, S. B., Harnisch, D. L., & Hastings, C. N. (1982). The validity of the Title I evaluation and reporting system. In E. R. House, S. Mathison, J. Pearsol, & H. Preskill (Eds.), *Evaluation studies review annual* (Vol. 7, pp. 427-442). Beverly Hills, CA: Sage Publications.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice*, *9*(3), 5-14.

Madaus, G. F. (1985). Public policy and the testing profession: You've never had it so good? *Educational Measurement: Issues and Practice*, *4*(4), 5-11.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. Champaign, IL: Stipes Publishing Co.

Meyer, R. H. (1996). Comments on chapters two, three, and four. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 137-145). Washington, DC: The Brookings Institution.

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

Olson, L. (1998, April 15). An "A" or a "D": State rankings differ widely. *Education Week*, *17*, 1, 18.

Raimi, R. A., & Braden, L. S. (1998). *State mathematics standards: An appraisal of science standards in 46 states, the District of Columbia, and Japan*. Washington, DC: Thomas B. Fordham Foundation.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1-7.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

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.

Richards, C. E., & Sheu, T. M. (1992). The South Carolina school incentive reward program: A policy analysis. *Economics of Education Review*, *11*, 71-86.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVASS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299-311.

Schmidt, W. H., & McKnight, C. C. (1998). *Facing the consequences: Using TIMSS for a closer look at US mathematics and science education*. Dordrecht: Kluwer Academic.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15-22.

Shepard, L. A., Smith, M. L., & Marion, S. F. (1996). Failed evidence on grade retention. *Psychology in the Schools*, *33*, 251-261).

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington DC: National Academy Press.

Tallmadge, G. K. & Wood, C. T. (1981). *User's guide: ESEA Title I evaluation and reporting system.* Mountain View, CA: RMC Corporation.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, *26*, 209-232.