The Design and Evaluation of Educational Assessment and Accountability Systems

CSE Technical Report 539

Robert L. Linn
CRESST / University of Colorado at Boulder

April 2001

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 2.2 Comprehensive Systems for Accountability and the Measurement of Progress, Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported 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.

THE DESIGN AND EVALUATION OF EDUCATIONAL ASSESSMENT AND ACCOUNTABILITY SYSTEMS

Robert L. Linn
CRESST/University of Colorado at Boulder

Abstract

Almost every state has in place a state assessment and accountability system. These systems vary greatly in their characteristics, but share a common global purpose of improving teaching and learning. Some of the variations in the state systems are discussed and illustrated with examples from selected states. Issues that are critical to the value and interpretation of results such as the use, if any, of comparisons among schools that serve students who come from different socioeconomic backgrounds, the relative weight given to current status or to improvement, and the basis for judging improvements at the school level (i.e., cross-sectional comparisons, quasi-longitudinal, and true longitudinal designs) are compared. The importance of evaluating and reporting the precision of assessment and accountability results is discussed. Finally, a key validity issue—the degree to which reports of performance and of improvement based on observed assessment results support inferences about student learning—is addressed. Evaluations of the degree of generalizability of results and trends through comparisons to other indicators of achievement and of improvement such as NAEP are stressed.

The current landscape of educational assessment and accountability systems is quite varied. At last count, every state except Iowa had adopted content standards, and most states have put in place assessment systems that arguably align to varying degrees with the adopted content standards. Iowa, the exception in terms of content standards, of course, also has a long tradition of testing. Almost all districts in Iowa have administered the Iowa Tests of Basic Skills (ITBS) each fall for the past several decades, but they do so on a voluntary basis rather than as a result of a state mandate.

The state systems that are in place differ along many dimensions including, but not limited to, those described here. They differ in terms of the uses that are made of test scores, the stakes that are attached to results for teachers, other educators, and students, the ways in which results are reported in terms of performance standards, how much emphasis is given to current status and how much to improvement, the

grades tested, the subject areas tested, the use, if any, of performance-based assessment tasks, whether normative comparisons are made, whether student socioeconomic status is taken into account, and whether students are tracked longitudinally. The states also vary a good deal with regard to the length of time that the assessment and accountability systems have been in place, the stability of the systems, and whether there are plans for phasing in new testing and accountability requirements over the next several years. Each of these dimensions has important implications for the design and evaluation of assessment and accountability systems.

Although it is commonplace to talk about education in the United States as 50 separate experiments, such a characterization suggests far more systemization, planning and evaluation than exists. Certainly, this is the case when it comes to the variations in state assessment and accountability systems. Although there clearly are rationales for the variations adopted by different states, any two state systems differ in terms of multiple factors that would make it difficult to attribute differences in effectiveness to any single factor, even if there were an agreed-upon basis for evaluating the effectiveness of the systems, which, of course, there is not. Consequently, we have learned less from observing variations in state systems than would be hoped.

We need a better basis for designing, evaluating, and redesigning assessment and accountability systems than we currently have. Since global comparisons of the effectiveness of systems are problematic, the best that can be done at this stage is to consider various design options one factor at a time and describe some of the pros and cons of different alternatives for each factor. We will attempt to do this by organizing the discussion around a series of questions.

Purpose and Intended Uses

The first question that should be asked of any assessment and accountability system is what are the purposes of the system? At a general level, the main purpose for all states is to improve instruction and student learning. The mechanisms that are assumed to help achieve this shared purpose vary a good deal. Tests and assessments are designed to provide information about student achievement. Providing such information to teachers, school administrators, parents, and students is expected to be helpful, although it is seldom specified in exactly what way. Teachers already have a great deal of information about student performance

through their day-to-day interactions with and observations of students, as well as their own teacher-constructed tests. State- or district-mandated tests provide an external check against which teacher judgments can be compared. Assessments that are aligned with state content standards help make the standards more explicit to teachers and students and may provide useful models for teaching and learning.

For external assessment to be really useful, teachers argue, such assessments need to provide diagnostic information; but external assessments are more suitable for providing global information about achievement than they are the kind of detailed information that is required for diagnostic purposes. Furthermore, the time frame for reporting results of external assessments is incompatible with the requirements of using diagnostic information to adapt instruction to student needs on a day-to-day basis.

Feedback provided to parents and students by external assessments is useful as a benchmark against which teacher reports of performance can be compared. A key rationale for the proposed Voluntary National Test (VNT) was that information from the VNT would provide parents with a basis for evaluating the achievement of their children against national standards of performance. The presumption was that parents could use that information to demand educational improvement if the performance of their children was found wanting.

For students, it is often assumed that tests, whether teacher constructed or externally mandated, can serve to motivate students to put forth greater effort to learn. Tests can focus attention on content subdomains and make clear performance expectations. This is true for teachers as well as students. Focus is desirable when there are concepts or problem-solving skills that are clearly important for students to master. There are tradeoffs, however. The sharper and narrower the focus, the easier it is to target teacher and student effort, but the costs may be more limited ability to generalize to new situations or problems and the neglect of other content areas. As will be discussed below, issues of teaching to the test, teaching the test, and cheating also arise.

In addition to providing information to educators, parents, and students, most state-mandated assessments have some type of stakes attached to the results. Even if there are no formally specified stakes, the results have stakes simply as a consequence of reporting aggregate results for schools to school boards and the public. Newspaper reports of results by school building create pressure on

principals and teachers to improve scores. Another source of pressure is the well-known use of test results for schools by realtors to entice prospective buyers. With the existence of the World Wide Web, access to information about the performance of students in a school is now readily available to interested parties on demand in many states. School report cards posted on the Web are now commonplace. From my home in Colorado, for example, I was able to find, with little time or effort, that less than a quarter of all third-grade students tested at a selected elementary school in Chicago scored high enough on the Illinois Standards Achievement Tests (ISAT) in Reading to meet the state standards, whereas two thirds of all tested third-grade students in an elementary school in Urbana, Illinois, met the state standards.

In addition to any stakes created by the public reporting of results, a plethora of types of stakes have been formally attached to assessment results. At the school level, these may involve accreditation requirements or the assignment of rewards and sanctions. For teachers they may involve monetary rewards in the form of bonuses or, in some instances, be the basis for pay-for-performance schemes. Negative consequences for teachers most commonly are informal ones, such as pressure from principals, but may include more formal actions, such as being singled out for some kind of assistance program. For individual students, accountability may involve placement in remedial programs, mandatory attendance of summer school, grade-to-grade promotion, or requirements for certificates of mastery, high school graduation, or level of endorsement on a high school diploma.

A fundamental premise of high-stakes accountability systems is that instruction and student learning will be improved by holding teachers and/or students accountable for results. This premise has broad popular appeal, but the empirical evidence regarding its veracity is mixed. There is a good deal of evidence that test scores generally increase during the first few years after a new assessment and accountability system is introduced. There is debate, however, about the degree to which the gains reflect real improvement or merely inflated test scores (see, for example, Linn, 2000). There is also disagreement about the prevalence and severity of unintended negative consequences of high-stakes assessments. It is clear, however, that high-stakes accountability systems are a prominent part of the educational agendas in states and districts throughout the nation.

Given the prevalence and demand for high-stakes accountability systems, it is clearly important that such systems be designed or redesigned so that they yield results that are valid, reliable, and fair. That is, there needs to be evidence that

supports the uses and interpretations of assessment scores. The scores need to have adequate precision for the decisions that are based on them. And, the system should not place some students, some teachers, or some schools at a relative disadvantage in comparison to others. Although the reliability and fairness criteria can be subsumed logically under the criterion of validity, they are important enough components of an overall validity evaluation that they deserve separate consideration.

Assessments

What should be assessed and how should it be assessed? The identification of purposes and intended uses of assessment results obviously has implications for what needs to be measured with regard to both the content areas and the nature of the assessments. Content standards that have been adopted by states are intended to specify what teachers are supposed to teach and what students are expected to learn. Although content standards are sometimes confused with a curriculum, the two are distinct. Content standards identify important concepts and skills that students are expected to learn, but they do not mandate a particular curriculum, textbook, instructional approach, or series of lessons. Content standards may serve as a guide for designing or evaluating curriculum, assessments, and instructional programs, but in each case, the intent of the standards could be met in a variety of ways.

Standards are expected to specify what should be taught and what students should learn. Assessments make those expectations concrete. They turn the statements about what students should know and be able to do into action. They provide the basis on which students and educators may be held accountable. Assessments are powerful tools in the use of standards to promote educational reform and improvement. Indeed, the nature of assessments and the ways in which they are used can determine the success or failure of the effort.

The power of assessments is due, in part, to their use in making the intended learning goals explicit. To do this it is essential that the assessments be aligned with the content standards. Alignment refers to the degree to which assessments adequately reflect standards. When closely aligned with the standards, assessments can reinforce the intent and priorities articulated in the standards. When poorly aligned, they can undermine and distort the standards. The expectations of the standards may be little more than hollow words, for example, if standards

emphasize the development of problem-solving skills and the application of those skills in everyday settings outside the classroom, while assessments require only the recall of simple facts and the recognition of right answers on frequently practiced problems. At the other extreme, important basic skills and core knowledge may be short-changed on an assessment that stresses only higher order conceptual understanding. Smith, O'Day, and Cohen (1990) emphasized the importance of alignment of assessments (examinations) with content standards (curriculum frameworks), noting that "the first and central lesson is this: If exams are used to motivate students to be more serious about their studies, then examinations' content must be closely tied to the curriculum frameworks that are used to teach students" (p. 41).

Assessments not only focus attention to the specifics of content and processes identified by content standards within a subject matter domain, they also reinforce the subjects that are assessed. Some subjects are privileged and are implied to be more important than others when assessments are administered for some content areas (e.g., English language arts and mathematics) but not others (e.g., science and history). This is clearly recognized by proponents of different subject matter areas who promote not only the development of content standards for their subject, but the inclusion of their subject as one of the subjects to be assessed.

Assessments need to be worthwhile targets for instruction and learning that not only are aligned with content standards, but communicate clearly the intent of the standards, encourage constructive action on the part of teachers and students, and are sensitive to instruction. This is a demanding set of expectations for assessments. Moreover, they need to be valid, reliable, and fair. These desired characteristics of assessments will be elaborated below in the context of discussing their implications for accountability systems.

School-Building Accountability

One widespread use of state-mandated assessments is for purposes of school-building accountability. Once it is decided to create a school-building accountability system, the question arises: How much emphasis should be given to current performance and how much to improvement? The most common way of reporting school-building assessment results is in terms of current status. This may be done by reporting the school mean or median score for students in the grade assessed using a scale score or a percentile rank metric or, as has become more popular in recent

years, the percentage of students who meet or exceed a performance standard or the percentage of students in each of several performance categories. Meyers (2000) has recently provided a critique of aggregate school-building indicators based on current status on a student assessment. He argued that such indicators, be they mean or median test scores, or a proficiency-level indicator, are "contaminated by factors other than school performance, in particular, the average level of achievement prior to entering first grade—average effects of student, family, and community characteristics on student achievement growth from first grade through the grade in which students are tested" (Meyers, 2000, p. 2). Meyers' critique of current-status school-building indicators discusses three additional shortcomings, but his first criticism alone is enough to raise serious questions about exclusive reliance on current-status school-building indicators to evaluate school performance, because, if used in isolation, they are unfair and will lead to invalid judgments regarding school quality.

Proponents of current-status indicators note that it is important to have the same high expectations for all children. The standards movement gives high priority to setting high standards of achievement for all students. Although it is recognized by even the strongest proponents of the standards movement that one cannot expect all students to meet standards overnight, and therefore, not all schools can be expected to achieve a desired level on a current-status indicator at this time, having all students meet standards remains a goal for some future date. Current-status reports are considered important because they reveal where students and schools stand at any given point in time, and when compared to desired performance targets, how far there is to go. The Florida school accountability system, for example, grades schools from A to F based on current performance of students on the Florida Comprehensive Assessment Test (FCAT). The purpose of the reports is described as follows. "The School Accountability Report groups schools with similar performance characteristics. It identifies critically low schools, stimulates academic improvement and summarizes information about school achievement, learning environment and student characteristics" (Florida Department of Education, 1999, p. 1).

Most state accountability systems that report school-building current status based on aggregate student assessment results also include some basis for rating improvement in achievement. This may be by comparing grade-level cohorts in the school in a given year (or years) with a cohort at the same grade for a previous year

(or years). It may involve comparing the performance of this year's fifth graders with that of students in the school who were in the fourth grade the previous year. Or, it may involve the comparisons of performance of students to the performance of those same students at an earlier point in time using matched longitudinal student records. Carlson (2000) referred to these approaches as cross-sectional, quasi-longitudinal, and longitudinal, respectively, and has presented analyses showing that they do not give the same answers to the question of which schools have shown the most improvement.

A number of the states also report a target performance level that all schools are expected to obtain by some specified date in the future. Colorado, for example, reports the percentage of students in a school who score at the proficient or advanced levels on their assessments and has set a target of 80% for schools to be accredited. There is also a provision, however, for schools with percentages below that level to be accredited if there is a 25% increase over the base-line percentage in a three-year period. California summarizes student performance using a scale that ranges from a low of 200 to a high of 1000 called the Academic Performance Index (API). A target of 800 that schools are expected to work toward has been set by the state. Annual growth targets on the API have also been set for schools.

Some states have fairly elaborate systems of grading schools in terms of current status and in terms of improvement. For example, as was previously noted, Florida assigns grades of A through F to schools based on the performance of the schools' students on the FCAT. The basis for assigning grades to schools is summarized in Table 1.

As can be seen in Table 1, grades of C through F are determined solely by student performance during the current year, whereas grades of A or B have added requirements for year-to-year change and requirements for the performance of subgroups of students.

The minimum and "higher" performing criteria referred to in Table 1 are defined by school level and content area in Table 2. As shown in Table 2, the minimum and higher criteria are defined by percentages of students at specified performance levels on the assessments in each of the three content areas. The state averages reported for 1999 FCAT indicated that, depending on school level and subject, between 70% and 78% of the students not exempted due to limited English proficiency or a learning disability scored at level 2 or above, and between 33% and

Table 1

Rules for Assigning Grades to Schools in Florida

Grade	
A	Meet grade "B" criteria AND the percent of students absent more than 20 days, percent suspended and dropout rate (high schools) are below state average AND there is substantial improvement[1] in reading AND there is no substantial decline[2] in writing and math AND at least 95% of standard curriculum[3] students were tested
B	Current year reading, writing, and math data are at or above higher performing criteria AND no subgroup[4] data are below minimum criteria, AND at least 90% of standard curriculum students were tested.
C	Current year reading, writing, and math data are at or above minimum criteria.
D	Current year reading, or writing, or math data are below minimum criteria.
F	Current year reading, writing, and math data are below minimum criteria.

Note. Grade description criteria and footnote quoted from Florida Department of Education (1999, pp. 1-2).

[1] Substantial improvement in reading means more than two percentage point increase in students scoring in FCAT levels 3 and above. If the school has 75% or more students scoring at or above FCAT level 3 AND not more than two percentage points decrease from the previous year then substantial improvement is waived.

[2] Substantial decline means five or more percentage points decline in the percent of students scoring FCAT achievement Level 3 and above OR five or more percentage points decline in the percent of students scoring 3 or above Florida Writes.

[3] Standard curriculum students also include Language Impaired, Speech Impaired, Gifted, Hospital Homebound and LEP students who have been in ESOL program for more than two years.

[4] Under current rule subgroups include economically disadvantaged, Black, White, Hispanic, Asian, and American Indian students.

Table 2

Criteria for School Performance Grades

Minimum criteria for school performance Grades C, D, and F				Higher performing criteria for school performance Grades B and A			
	FCAT Reading	FCAT Math	Florida Writes!		FCAT Reading	FCAT Math	Florida Writes!
Elementary	60% score level 2 & above	60% score level 2 & above	50% score 3 & above	Elementary	50% score level 3 & above	50% score level 3 & above	67% score 3 & above
Middle	60% score level 2 & above	60% score level 2 & above	67% score 3 & above	Middle	50% score level 3 & above	50% score level 3 & above	75% score 3 & above
High	60% score level 2 & above	60% score level 2 & above	75% score 3 & above	High	50% score level 3 & above	50% score level 3 & above	80% score 3 & above

Note. Grade description criteria quoted from Florida Department of Education (1999, p. 1).

51% of the students scored at level 3 or above. Thus, the criteria for a school to receive a grade of C are reasonably demanding whereas those for a grade of A or B are a good deal higher than can be met by the typical performance of students for the state as a whole.

A number of other states have rules for grading schools that are at least as complicated as those illustrated above for Florida. The ways of grading improvement are also quite varied and often complicated to explain. Massachusetts, for example, has a system that uses performance ratings based on current status of student achievement on the Massachusetts Comprehensive Assessment System (MCAS) not only to place schools in performance categories, but also to determine targets for improvement. Schools are placed in one of six performance categories based on the percentages of students in a school scoring in the proficient and advanced ranges and the percentages with "failing" scores on the MCAS. Improvement over a two-year cycle is then measured in terms of increases in the school's average MCAS scaled score for each content area. The minimum amount of improvement in average scaled score that schools are expected to achieve is then set for each of the six performance categories, as is shown in Table 3.

Based on criteria for improvement, specified in Table 3, schools are given one of three improvement ratings: "Failed to Meet," if the average scale score improvement is more than 1 point below the target increase; "Approached," if the

Table 3

Massachusetts MCAS Performance Categories and Improvement Expectations

Performance categories	Percentage of students scoring in proficient or advanced		Percentage of students scoring failing level	Increase average scaled score by:
1	80% or more	and	5% or less	1-3 points
2	60% or more	and	10% or less	1-3 points
3	40% or more	and	20% or less	2-4 points
4	20% or more	and	40% or less	3-5 points
5	Less than 20%	or	60% or less	4-6 points
6	More than 60%			5-7 points

Note. From Massachusetts Department of Education (1999, p. 3).

improvement is within 1 point of the target increase; or "Met," if the improvement is at or higher than the lower bound of the target increase range (Massachusetts Department of Education, 1999, p. 3).

A couple of features of the Massachusetts school-building accountability system are worthy of note. First, the target increases are higher for low-performing schools than for high-performing schools. Second, and of greater significance, improvement is a secondary consideration in the accountability system. Although the three improvement categories provide schools in low performance categories with some potential consolation, it is the overall performance category based on current status summarized over two school years that leads to labeling schools.

"An overall performance rating for school will be calculated by averaging across the content areas the percentage of students scoring in the Failing and Proficient or Advanced levels on MCAS tests administered during the two-year rating cycle. The performance category into which the school's two-year average falls will determine the school's overall performance rating. From the highest to lowest performance categories listed in . . . [Table 4], overall performance ratings will be as follows: Very High, High, Moderate, Low, Very Low, and Critically Low" (Massachusetts Department of Education, 1999, p. 3). Given Meyers' (2000) critique of current status as a means of school-building accountability that was previously quoted, it is clear that schools serving large numbers of poor students do not have a fair chance of achieving one of the higher performance categories.

Like Massachusetts, Kentucky also has a system that sets higher targets for improvement for schools where student achievement is low over a two-year baseline. Unlike Massachusetts, however, Kentucky places the primary emphasis on improvement rather than current status. Using a scoring scheme that assigns a value of 100 to assessment scores in the proficient range, a value higher than 100 to scores in the advanced achievement range, and values less than 100 for scores below the proficient range, a long-range target of 100 was set for all schools. The gains the schools are expected to achieve in each biennium are set such that all schools continuing to meet their targets would achieve 100 at some specified date in the future. Schools are placed into accountability categories and given rewards or identified as in need of assistance not on the basis of current-status performance scores, but on the basis of where they stand with respect to the targeted improvement.

Socioeconomic Background

Should socioeconomic factors be taken into account? It is well known that socioeconomic background is substantially related to student achievement. But the existence of a relationship does not lead to an obvious choice of whether or not socioeconomic factors should be taken into account before passing out rewards and sanctions to schools based on student achievement. Elmore, Abelmann, and Fuhrman (1996) characterized the issue as follows: "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).

Different states come out on different sides of the issue of making adjustments for socioeconomic status (SES). Pennsylvania, for example, uses a number of community type and SES variables to identify similar schools ("10 schools scoring immediately below and 10 schools scoring above the target school"; Pennsylvania Department of Education, n.d., p. 24) and then reports the interquartile range for reading and mathematics scores for the set of similar schools called the "Similar Schools Score Band." The Pennsylvania Department of Education explains the reasons for using similar school bands to report results as follows: "It is well established that academic achievement is influenced primarily by two factors: the quality of the educational services provided and the socioeconomic backgrounds of the students themselves. These factors might be classified as 'school' and 'non-school' factors. Similar school information permits a school to compare its results with those of the same community type and socioeconomic background. The Similar Schools Score Band, therefore, supplements the comparison of school score with the overall state average" (Pennsylvania Department of Education, n.d., p. 24).

The use of comparisons to other schools with similar SES characteristics is not unique to Pennsylvania. California is one of several other states that use similar schools defined by SES factors as one basis of comparison within an overall system of school-building accountability (California Department of Education, 2000). In most cases, the bands of the similar school results are a secondary consideration that provides another basis for judging results in addition to the main accountability results that do not take SES into account. The use of SES to make adjustments or as

the primary basis of accountability is problematic because of the concern quoted above from Elmore, Abelmann, and Fuhrman (1996) regarding the institutionalization of different expectations for different groups of students. The reason that SES adjustments are problematic is summarized concisely by Clotfelter and Ladd (1996) as follows: "If one uses socioeconomic status as a predictor, the effect is to set a lower threshold for success for poor students than for rich ones" (p. 26). The problematic nature of SES adjustments is exacerbated by the fact that there is a strong relationship between SES and ethnicity. Consequently, lower standards for students from low SES backgrounds automatically mean lower standards for African American and Hispanic students because of the relationship between SES and ethnicity.

Prior Achievement

Should prior achievement be taken into account? Accountability systems that emphasize change in performance over time rather than current status provide a means of taking into account characteristics of the students attending the school without resorting to measures of SES, which are at best only indirect proxy measures of the level of achievement that students bring to school at the start of any grade level. Systems such as those in California, Colorado, Kentucky, Maryland, and Washington that compare the achievement of students at selected grades in a given year or biennium with that of cohorts of students from previous years at the same grade in the same school provide a means of recognizing that schools serve students that start at different places. Such cross-sectional comparison of students at a grade level in different years rests on the implicit assumption that the student characteristics that affect initial achievement levels are relatively stable from year to year for the students attending a given school. This assumption is questionable for schools serving neighborhoods whose demographic characteristics are changing rapidly, but it is reasonable in a rough sense for most schools.

A more direct way of taking prior achievement of students into account is, of course, to track changes in student achievement from one grade to the next. There are two ways in which this is done in state accountability systems. The first is to simply compare the achievement of students in one grade in a prior year with that of students in the next higher grade the following year. Such an approach is what Carlson (2000) has recently called a quasi-longitudinal analysis. It has the advantage of simplicity over a true longitudinal design, which tracks the achievement of students with matched records from one year to the next. The quasi-longitudinal

approach also has the advantage of including all tested students each year, not just those who remain in a school and have matched assessment results for both years or, for some analyses, for multiple years. Using gains in achievement from one grade to the next for either a quasi-longitudinal or true longitudinal analysis requires that the assessment results be reported on what is known as a vertical scale. That is, the scores reported for fourth-, fifth-, or sixth-grade students, or any other combinations of grades, need to share a common metric despite the fact that students in different grades are administered different assessment tasks.

North Carolina is an example of a state that uses a quasi-longitudinal approach in its "ABC" school-building accountability system. "The ABCs of Public Education is a comprehensive plan to recognize public schools in North Carolina. This plan focuses on (1) strong accountability, (2) emphasis on the basics and on high educational standards, and (3) maximum local control. A key component of the ABCs of Public Education is a new accountability program, which focuses on performance of individual public schools (rather than school systems) in the basics of reading, writing, and mathematics. Rather than comparing different students from one year to the next, this plan-the-school-based Management and Accountability Program holds schools accountable for the educational growth of the same groups of students (cohorts) over time. At least a year's worth of growth for a year's school is expected" (North Carolina Department of Public Instruction, 1996, p. 1).

North Carolina uses the average rate of growth observed across the state as a whole from one grade in the spring of 1993 to the next grade in the spring of 1994 as a benchmark against which the improvement for students in a given grade in one year to the next grade the following year is judged. The details of how school changes in achievement in, say, third grade in 1999 to fourth grade in 2000, are evaluated is complicated in that allowances are made both for differential expected rates of growth for students at different points on the scale and regression to the mean effects, but the basic idea of the system is straightforward. The 93-94 state average growth figures set an expectation for the school-building growth for a given pair of grades and a given pair of years of assessment after the school-building results have been adjusted for differential growth rates and differential regression effects. The comparisons to expected growth are then used to classify schools into one of four categories: exemplary schools, schools meeting expected growth, schools having adequate performance, and low performing schools.

The Tennessee Value-Added Assessment System (TVAAS) is perhaps the best known and most often cited state accountability system that relies on matched student-level longitudinal data for reporting of school, district, and teacher performance. The TVAAS reports, like those for North Carolina, use a vertical scale to report student test results. As noted by Bock and Wolfe (1996) the common scale is essential for the computation of "gain" scores used in the TVAAS analysis and reporting procedures. TVAAS was developed by William L. Sanders (see, for example, Sanders & Horn, 1994; Sanders, Saxton, & Horn, 1997) using sophisticated data analysis methodology that allows the use of gains in student achievement from one year to another as the basis for holding teachers, schools and districts accountable. Student achievement data from several previous years are used as the basis for estimating gains in a particular year. The statistical model allows for missing data, so a student with scores missing for some of the prior years, but who has data for at least one of the prior years can be included in the analysis of gains for a given year. District-level, school-level, and teacher-level contributions to student gains are estimated. Each of these contributions is interpreted as the value added by a teacher, a school, or a district.

An elaborate longitudinal database has been developed that allows TVAAS not only to track individual students' test performance over several years, but to associate each student's scores with the teachers he or she had in each grade for which test results are available in the database. This has enabled Sanders and his colleagues to investigate growth patterns for students who have teachers who are estimated to make positive contributions to student test scores for several years in a row, or who have various patterns of teachers with estimated contributions to student test score of different magnitudes. Sanders and Rivers (1996), for example, reported that students who had teachers estimated to make high contributions to student test score gains had fifth-grade math scores that were slightly more than 50 percentile points higher than scores of students who started at similar levels but had teachers estimated to make low contributions to student gains for three years in a row. Students with various combinations of high, average, and low teachers over the three years had average fifth-grade math scores that fell between those extremes.

Using prior achievement of students as a predictive factor in an accountability system has many advantages over systems that rely on SES factors to adjust scores or to produce comparison bands of schools. Unlike adjustments for SES, the use of prior achievement as predictor of subsequent achievement does not establish

different gains for students from different backgrounds. Since prior achievement is substantially correlated with student SES, it provides a basis for taking into account a large part of the SES differences that proponents of using SES seek to take into account to provide fair comparisons that depend only on factors that teachers and schools can influence rather than differences in student cohorts served. There is a question, however, whether taking prior achievement into account completely levels the playing field or whether there remains some residue of differences in SES factors that still advantage some teachers and schools over others.

Results reported by Sanders and his colleagues suggest that taking prior achievement into account as it is done in TVAAS is all that is necessary to yield a fair basis of comparison. In a document reporting frequently asked questions and answers, for example, Sanders and Horn (n.d.) give the following question and answer. “My students are mostly from the inner city. Won’t that make a difference in their gain scores?” Answer: “The pilot studies revealed no relationship between the racial composition of student body and gain scores. Whether a school was an inner city school or a suburban one was also found to be unrelated to gains made” (p. 5 after title page; pages of the document are unnumbered). This general conclusion was reaffirmed by Sanders and Rivers (1996).

Recently reported results by Hu (2000) call into question the conclusions by Sanders and his colleagues that race/ethnicity and SES of the student body are unrelated to the gains estimated in TVAAS. Hu obtained school-building data on per pupil instructional expenditures, the percent of minority students in the student population, and the percent of students eligible for free or reduced-price lunch. He correlated these variables with value-added estimates based on three-year averages across grades for reading and mathematics. For the 58 elementary schools in his study, Hu found that per pupil instructional expenditures had correlations of .39 with the average value added in both mathematics and reading. The correlations for percent minority were .42 and .28 for mathematics and reading, respectively. The corresponding correlations for percent free or reduced-price lunch were .49 and .29. The squared multiple correlations of all three school factors with the three-year averages of value-added estimates from TVAAS were .27 for reading, .19 for mathematics, and .28 for the composite of reading and mathematics. Thus, between a fifth and a bit more than a fourth of the variability in the value-added three-year averages was predictable from a combination of per pupil instructional expenditure,

percent of minority students in the student body, and percent of students eligible for free or reduced-price lunch.

Hu's findings lend support to the observation by Shepard, Kupermintz, and Linn (2000) that although TVAAS adjusts for differences in student achievement it does so imperfectly. Relationships of TVAAS gains with variables such as percent of minority students in the student body and the percent of students eligible for free or reduced-price lunch are consistent with the notion that the adjustments are imperfect. Adjustments for differences in student achievement do not preclude the possibility that students from different SES backgrounds will have different levels of support for learning and differential access to enrichment experiences outside of school during the year in which gains are being estimated and thereby yield systematic biases in the estimated school and teacher effects. Although the adjustments for differences in student achievement go a long way toward leveling the playing field, they may fall short of fully accomplishing that end.

Another fundamental criticism of the TVAAS model, as well as other longitudinal and quasi-longitudinal models that depend on annual testing in every grade, that is discussed by Shepard, Kupermintz, and Linn (2000) has to do with limitations in the assessments used. Because of the requirements of annual testing of students in every grade and the need to have a common vertical scale for reporting results, there is a tendency to use publisher-provided standardized tests that are either off-the-shelf tests or ones that are highly similar in their characteristics to off-the-shelf tests. Such tests are almost sure to be less well aligned with state content standards than an assessment that is specifically designed to measure the knowledge and skills emphasized in the content standards. The same test forms also tend to be reused.

As was argued above, assessments need to be closely aligned with content standards and to communicate the intent of the standards in order to provide worthwhile targets for instruction and learning. Tests that are out of alignment can undermine the intent of the standards and distort instruction to match the tests rather than the standards. Reuse of the same test forms year after year creates even more serious problems, not only distorting instruction, but also narrowing it to the specifics of the test. It is worth repeating in this regard an observation made by Cronbach (1963) nearly four decades ago: "Whenever it is critically important to master certain content, the knowledge that it will be tested produces a desirable concentration of effort. On the other hand, learning the answers to a set of questions

is by no means the same as acquiring understanding of whatever topic the question represents" (p. 681). Teaching *the* test undermines the ability to generalize. Even teaching *to* the test can limit generalizations if the same form of the test is used from year to year. And, the goal of generalizing to the content standards can only be achieved if the assessments are aligned with the standards and change the samples of aligned tasks that are included from year to year.

Technical Quality of Assessments Used for Accountability

How should the technical quality of assessments used in high-stakes accountability systems be evaluated? There is a broad professional consensus in the educational measurement community that the requirements for technical quality and evidence supporting the uses and interpretations of high-stakes assessments are more stringent than they are for assessments with low stakes. The relevance of stakes in evaluating an assessment system is made clear in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999; hereafter the *Standards*). The *Standards* are the most authoritative statement of professional consensus of standards for testing practice available. They provide broad guidance for those attempting to design or evaluate assessment programs and accountability systems that make use of student assessment results. Although the principles articulated in the *Standards* have general relevance for low-stakes uses of assessment results such as the diagnostic uses of test data by teachers, such uses raise relatively few questions about validity or other technical qualities, in part, because mistakes can be quickly corrected in light of new information and, in part, because the decisions do not have major consequences for students. High-stakes uses of test scores, on the other hand, are expected to meet higher standards and to provide evidence to support claims of reliability, validity, and fairness.

Reliability

Reliability refers to the precision of assessment scores and is usually gauged in terms of the consistency of scores obtained using alternate sets of assessment tasks, or different occasions, or when different raters score open-ended responses. All assessments are fallible; that is, they have less than perfect precision and produce less than perfectly consistent results. The task of reliability analyses is to quantify the degree of precision or its converse, the degree of impression. The *Standards* note that "precision and consistency in a measure are always desirable. However, the need for

precision increases as the consequences of decisions and interpretations grow in importance" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999, p. 30).

There is a popular belief that assessments have greater precision than they actually have. This is due, in part, to the appearance of precision that is provided by a test score. The three-digit reports of SAT scores by the College Board, for example, make a score of 530 seem superior to one of 500, when, in fact, the two scores are well within the bounds of what would be expected as the result of measurement error. Similarly, a report that a student scored at the 70th percentile on a mathematics test makes it seem that the student is clearly above average, but as Rogosa (1999a) has shown, there is a substantial likelihood that such a percentile rank will be achieved by a student who is truly below average on tests that have reliability coefficients that are of respectable magnitude. When currently popular performance standards are used in score reporting, it rarely occurs to users of the scores that the student who scores in the "partially proficient" range may actually be "proficient" but simply scored in the partially proficient range due to measurement error. Analyses reported by Rogosa (1999a) reveal that the probabilities that students will be misclassified as the result of measurement error are considerably larger than is generally assumed.

Reporting reliability information. Reliability coefficients are the most common way of reporting information about precision or consistency of test scores. Although useful for certain purposes, reliability coefficients do not do a very good job of summarizing the degree of precision of assessment and often convey an exaggerated view of how precise the measurement is. As noted in the *Standards*, "the standard error of measurement is generally more relevant than the reliability coefficient once a measurement procedure has been adopted and interpretation of scores has become the user's primary concern" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999, p. 29). Unlike a reliability coefficient, a standard error of measurement provides information about the magnitude of error that is likely to be associated with scores in units of the scale used for reporting. Knowing, for example, that the standard error of measurement is roughly 30 points on the SAT score scale, makes it clear in the above example that a score of 530 differs from one of 500 by an amount that could easily just be the result of measurement error. This is not apparent from a reported reliability coefficient of .90.

A currently popular way of reporting results of state assessments is in terms of performance standards. The number of performance standards set for a grade and content area varies from state to state, but is generally greater than the two points dividing performance into three levels required by IASA. Kentucky, for example, has set three performance standards in each content area resulting in four levels of achievement, which are labeled distinguished, proficient, apprentice, and novice (Trimble, 1994). The National Assessment of Educational Progress (NAEP) also uses four categories of performance, which are labeled below basic, basic, proficient, and advanced. Maryland has set four standards that yield five levels of achievement for the Maryland School Performance Assessment Program (MSPAP). Even publishers of norm-referenced tests have joined in the trend to report results in terms of performance standards. Harcourt Educational Measurement, the publisher of the Stanford 9 (SAT9), for example, has set performance standards at each grade level to allow reporting in terms of advanced, proficient, basic, and below basic performance categories. Other publishers have also set performance standards for their tests. Whatever the number of performance levels distinguished, the central question about precision of the assessment is the probability of misclassification—for example, calling a student partially proficient who is really proficient.

The importance of evaluating the precision is recognized in the *Standards*. "When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999, Standard 2.15, p. 35).

Classification accuracy is dependent upon score reliability, on the number of performance levels, and on the location of the cut scores that distinguish between adjacent performance levels. For a given level of reliability, classification accuracy will decrease as the number of levels increases. It will also be lower when cut scores are closer together than when they are farther apart. Obviously, students whose true level of performance is near the cut score between two categories have a higher probability of being misclassified than students who are farther from the cut score. Rogosa's (1999a) analyses show that there are relatively high probabilities of misclassifying students whose true performance differs from the cut score by amounts equivalent to 10, 15, or even 20 percentile points. This is true even for tests

that have reliabilities that are considered high, say .85 or .90. For a test with a reliability of .90 where the proficient level was set at a score equal to the 50th percentile, a student whose true percentile rank was 30 would have a probability of .06 of scoring in the proficient range on the test. Students with true percentile ranks of 35, 40 and 45, who, like the student with a true percentile rank of 30, should be classified as below proficient, would have probabilities of .12, .22, and .35 of being misclassified as proficient due to errors of measurement. Of course, it is not a sure thing that students who deserve to be classified as proficient will be so classified. For example, students whose true percentile ranks are at the 55th, 60th, 65th, and 70th percentiles all deserve to be classified as proficient, but the probabilities that they will be misclassified in a below proficient category on a test with a reliability of .90 are .35, .22, .12, and .06, respectively (see Rogosa, 1999a, Exhibit IF, p. 29). If high-stakes decisions are based on the assessment, these probabilities of misclassification are disturbingly large.

Standard errors of measurement and classification accuracy are relevant for school-level results as well as for individual student results. The mean score of a school has a standard error of measurement associated with it that is due not only to the sampling of items by the assessment, but to the sample of students who take the test, although there is a perspective that argues that since all students in a grade are tested, it is a population result. However, treating the students as a fixed set implies that the results should be treated only as an historical fact rather than as an indication of school quality or a measure against which school progress will be judged by comparing the results to those for the same or a different cohort of students at a later point in time. For the latter interpretations, students need to be considered a sample, and sampling variability need to be considered to be a contributor to the measurement error of the school mean or other school-building statistics such as the percentage of students achieving at the proficient level or higher (for an elaboration of this argument, see Brennan, 1995, and Cronbach, Linn, Haertel, & Brennan, 1997).

School-building errors of measurement are heavily dependent on the number of students in the school at the grade assessed (see, for example, Burton, 2000; Linn & Burton, 1994). The implications of measurement error for school-building accountability systems depend on the rules of the accountability system, but regardless of the rules, it is clear that the likelihood of misclassifying a school will be greater for small schools than for large schools.

Analytical procedures such as those used in the Sanders value-added model produce estimates of standard errors. As Bock and Wolfe (1996) have shown, however, the model-based standard errors are underestimates because "they neglect the intraclass correlation in scores of students within the same teacher-classrooms" (p. 56). The empirical standard errors reported by Bock and Wolfe for three consecutive years of three-year school averages for the TVAAS model were between 56% and 108% larger than the model-based standard errors reported by TVAAS depending on the grade level and subject area tested. In the most extreme case of the social studies test at Grade 8, the empirical standard error was almost as large as the between-school standard deviation (4.08 vs. 4.27). Even in the best case of Grade 5 mathematics, the empirical standard error was 59% the size of the between-school standard deviation (5.00 vs. 8.48). Bock and Wolfe's (1996) review of the empirical standard errors in comparison to the between-school standard deviations led them to question the adequacy of the school-building reports. In their words, "the results raise the question whether it is advisable to publicly report these scores at their present level of accuracy. A more prudent course would be to examine the distribution of school gains for the state as a whole and look for additional evidence that the schools with extremely high or low gains are memorable in other ways that would explain their positions in the distribution" (p. 58).

Where schools are placed into categories as is done in several of the school-building accountability systems described above (e.g., Florida's A through F ratings and the six categories used by Massachusetts), attention needs to be given to the likelihood that schools will be misclassified as the result of measurement error. There are several ways of evaluating classification accuracy for school-building accountability systems. The key to all approaches is to fully model the ways in which the accountability indices are constructed and the boundaries used to classify schools.

Procedures for evaluating school-building misclassification probabilities are described by Rogosa (1999b). Rogosa's results show that the probabilities of misclassifying schools are nontrivial. Alternative approaches to evaluating the likelihood that school buildings will be misclassified are described by Hoffman and Wise (2000). Kentucky has a contract with HumRRO to evaluate the accuracy of the classification of school buildings for its accountability system using the analytical procedures described by Hoffman and Wise. Such investigations of the accuracy of

accountability system classifications of schools need to be a standard part of the evaluation of the technical adequacy of accountability systems.

Systems that use student assessment results as part of teacher evaluation systems also need to evaluate the precision of the teacher-level information, especially if there are stakes attached to the results such as mandatory assistance or pay for performance. TVAAS again provides an example. The TVAAS approach uses regressed estimates of gain scores associated with teachers. These estimates have the advantage that the greater uncertainty for teachers who have fewer students in the calculations have their gains pulled back closer to the overall mean. On the other hand, more reliance is placed on the teacher-specific results for teachers when there is greater certainty as the a consequence of having a larger number of students with data for the calculations. Nonetheless, it is important to evaluate the magnitude of the standard errors of the "teacher gain scores." Bock and Wolfe (1996) computed empirical standard errors for the teacher gain scores and found that they were generally larger, albeit by a relatively small amount, than the model-based estimates produced by TVAAS. They concluded that "although the estimates are . . . variable from year to year, the results were stable enough to permit identification of teachers with notably meritorious or problematic instructional effectiveness, as measured by test-score gain" (p. 71). They went on to recommend, however, that the results should be reported in ways that make the magnitude of the standard errors evident, for example, by graphical displays that show confidence intervals for the teacher gains. If this were done it, would make it obvious, as is evident from the example Bock and Wolfe provide on page 66, that some teachers with gains in the middle range may actually be indistinguishable from some other teachers with gains in the high or low categories.

Validity

Validity is the most fundamental consideration in the evaluation of the uses and interpretations of any assessment. But, how should validity be investigated and reported? Validity is such a broad concept that there is a need to get clear, as Shepard (1993) has suggested, about the questions that are of the highest priority to address. Since validity is specific to particular uses and interpretations, it clearly is not appropriate to make an unqualified statement that an assessment is valid. Rather, the assessment that has a high degree of validity for a particular use may have little or no validity if used in a quite different way. For this reason, the *Standards* admonish the developers and users of assessments to start by providing a

rationale “for each recommended interpretation and use” (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999, p. 17).

The earlier discussion of alignment is relevant to the evaluation of validity of assessments, the results of which are interpreted as measures of state content standards. The evaluation of the alignment of tests with content standards is often much too superficial. If asked whether their tests are aligned with the content standards of a state, any test publisher can be counted on to give an affirmative answer. But the answer is unlikely to stand up to close scrutiny. No test or assessment is likely to cover the full domain of a set of content standards. Even those aspects that are covered will vary in the degree and depth of coverage. Hence, an adequate evaluation of alignment must make it clear which aspects of the content standards are left uncovered by the test, which are covered only lightly, and which receive the greatest emphasis. Such an analysis provides a basis for judging the degree to which generalizations from the assessment to the broader domain of the content standards are defensible. If only aspects of the domain that are relatively easy to measure will be assessed, this, in turn, can lead to a narrowing and distortion of instructional priorities.

The use of off-the-shelf tests for high-stakes accountability often leads to practices that undermine the validity of inferences about the achievement domains that the tests are intended to assess. The use of “scoring high” materials closely tailored to particular standardized tests is designed to raise scores. But increased scores do not necessarily mean that improvements would generalize to a domain of content that is broader than the test.

Test preparation materials designed to help students score high on tests are not limited to off-the-shelf standardized tests. A January 1999 American Guidance Service (AGS) advertisement called *Taking the Terror Out of the ITBS,* for example, claims, “Test Scores Increase by Over 300% at Georgia Elementary School,” while another advertisement makes the promise: “Raise Your Test Scores 20 to 200%” (Evans Newton advertisement for Target Teach, January 1999). Test preparation materials for both teachers and parents are readily available online, not just for standardized tests, but for a number of state assessments as well. Sleek Software Corporation, for example, makes test preparation materials for the Texas Assessment of Academic Sills (TAAS) available to teachers and to parents at their Web site, where the materials are described as follows. “Sleek Software’s Incredible

Tutor (or IT! for short) is the most comprehensive TAAS preparation tool available anywhere. . . . The content is written specifically to match the format of the TAAS test . . . Comprehensive lessons and examples are provided" (http://www.sleek.com/TX/ITTX.html).

The boundary between acceptable forms of test preparation and cheating are fuzzy and confusing to the public and to many educators. Teachers naturally want to help students be prepared to take tests. When accountability includes stakes for teachers based on the performance of their students on tests, there are added incentives to help students achieve the best scores possible. Certainly it is legitimate—indeed desirable—to focus on areas emphasized in content standards and the curriculum that are apt also to be emphasized on the tests. It is legitimate to provide students with practice on the format and generic types of problems that they will encounter on the test.

Practice on items that are clones or slight variations on actual test questions is more problematic. Some testing experts consider it unethical to provide such practice (see, for example, Haladyna, Nolen, & Haas, 1991, and Mehrens & Kaminski, 1989, for more complete discussions of the issues). For others, it may simply be considered poor or impoverished instructional practice. Giving students advance exposure to the actual test questions or paraphrases of those questions is beyond the pale of legitimate practice and should be included with other even more blatant forms of cheating such as giving students clues or giving them the answers while they are taking the test or altering their answers. See Hoff (2000) for a recent discussion of the confusion and different interpretations of acceptable practice.

Generalizability of Gains in Scores

Gains in scores on state assessments are generally interpreted to mean that student achievement, and by implication, the quality of education, has improved. The reasonableness of such an interpretation depends on the degree to which generalizations beyond the specific assessment administered by the state to the broader domains of achievement defined by the content standards are justified. A variety of factors, such as teaching that is narrowly focused on the specifics of the assessments rather than on the content standards they are intended to measure, may undermine the validity of desired generalizations.

Gains in test scores were commonly observed on the norm-referenced tests widely used by states during the 1980s. Indeed, in what came to be known as the

Lake Wobegon effect (Koretz, 1988) almost all states were reporting results on norm-referenced tests that were above the national average defined by the test publisher norms. John Cannell called the Lake Wobegon effect to the attention of the public. Cannell (1987) accused states and test publishers with intentionally reporting results that were misleading the public. There were a variety of reasons other than the rather cynical ones emphasized by Cannell for the Lake Wobegon effect—for example, the use of old norms in a period when student achievement was generally increasing across the nation, reuse of the same form of the test year after year, and the exclusion of more students from testing on state administrations than in the publisher norming studies (for elaborations of these and other potential explanations, see Koretz, 1988; Linn, Graue, & Sanders, 1990; Shepard, 1990).

Other than the explanation of old norms, most of the reasons that led to the inflated scores on norm-referenced tests during the period of the Lake Wobegon effect remain potential concerns with the standards-based reporting that is currently prevalent for state assessment and accountability systems. Hence, it is important to evaluate the degree to which generalizations of gains on assessments to broader domains of achievement are justified. One practical and relatively powerful way of investigating generalizability is to compare trends for state assessments with trends for the state on the National Assessment of Educational Progress (NAEP). An example of such a comparison is shown in Figure 1: the percentage of fourth-grade students in Kentucky who scored at the apprentice level or higher in reading on the Kentucky Instructional Reporting and Information System (KIRIS) for the years 1993 through 1998. Also shown are the percentages of fourth-grade students in Kentucky who scored at the basic level or higher on NAEP for 1992, 1994, and 1998, the years when reading was assessed by NAEP at Grade 4 at the state level. Although there is no necessary correspondence between the apprentice level on KIRIS and the basic level on NAEP, it can be seen that the 58% of students at the basic level or higher on NAEP in 1992 is not far below the 68% of students at the apprentice level or higher on KIRIS in 1993. The increases in the percentage of students scoring at the apprentice level or higher on KIRIS in the following years, however, are not mirrored by comparable increases in the percent basic or higher on NAEP. Indeed, while the KIRIS percent increased from 68% from 1993 to 1994, the NAEP percent actually decreased slightly (from 58% to 56%). Between 1994 and 1998, the percent of

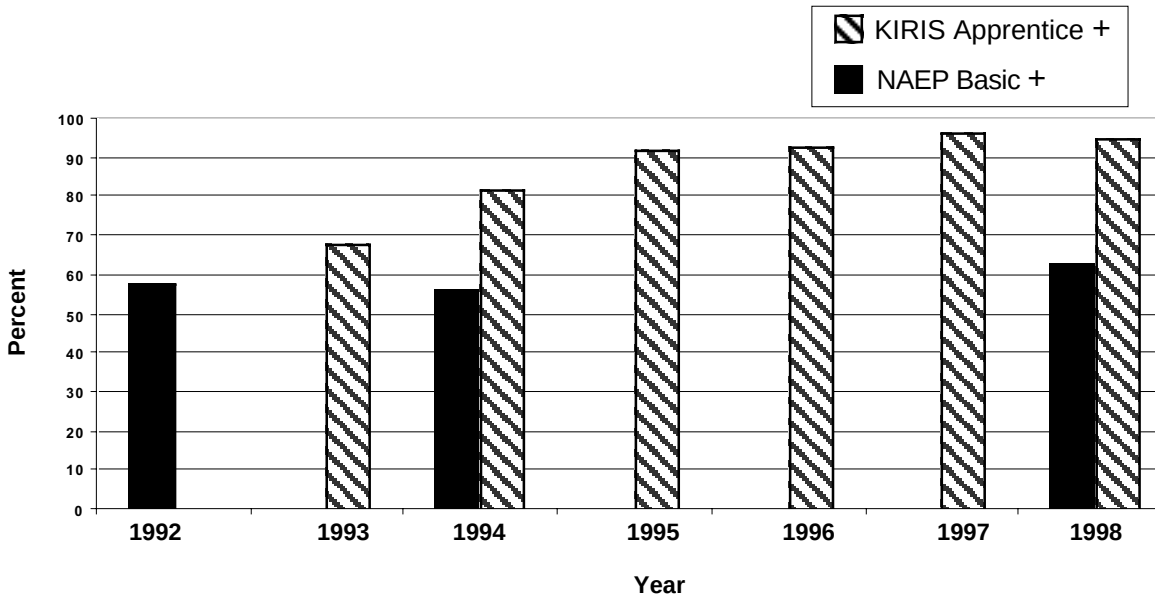

Figure 1. Kentucky fourth-grade reading trends (KIRIS percent Apprentice or Above vs. NAEP percent Basic or Above).

students who scored at the apprentice level or higher on the KIRIS continued to rise, reaching highs of 96% in 1997 and 95% in 1998. During the same period, the percent of students who scored at the basic level or higher on NAEP also increased, but by a more modest amount (from 56% in 1994 to 63% in 1998). A more detailed comparison of the gains on KIRIS with those on NAEP as well as on college admissions tests (the ACT) is provided by Koretz and Barron (1998). Their results reinforce the observation from the results displayed in Figure 1 that increases shown on KIRIS do not generalize very well to other indicators of student achievement such as NAEP. As I have argued elsewhere, "divergence in 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" (Linn, 2000, p. 14).

Klein, Hamilton, McCaffrey, and Stecher (2000) have recently reported the result of a series of comparative analyses of the trends for the Texas Assessment of Academic Skills (TAAS) and NAEP. They compared the trends for TAAS at Grades 4 and 8 in mathematics and Grade 4 in reading with those observed for the state of Texas on NAEP separately for White, African American, and Hispanic students. Their results show not only that the increases in scores for the three subgroups do not generalize to NAEP, but that the closing of the gap in performance between the

two minority groups and Whites that has been observed on TAAS is not replicated on NAEP. According to TAAS, the gap between White students and students of color in Texas decreased substantially between 1994 and 1998. According to NAEP, however, the gap actually increased slightly during this period. Given the pressure to improve scores on TAAS, especially in schools serving students of color who have had low scores on achievement tests in the past, the failure for the overall trends and the narrowing of the gap to generalize to NAEP results raises serious questions about the trustworthiness of the TAAS result for making inferences about improvements in achievement of students in Texas or about the relative size of the gains for different segments of the student population.

Summary and Conclusions

State assessment and accountability systems vary greatly along a number of dimensions, including the subjects that are assessed, the nature of the assessments that are used, the stakes that are attached to results, and whether those stakes are for students or educators, or both. They also vary in their reliance on current achievement results versus the emphasis that is placed on improvement and whether the system relies on cross-sectional, quasi-longitudinal, or true longitudinal data where individual students are tracked over time. Regardless of the details of the systems, all state assessment and accountability systems have the same global purpose: the improvement of instruction and student learning. There is considerable debate, however, over the degree to which the systems contribute to that goal.

There is widespread agreement that assessments play an important role in shaping instruction and thereby influencing student learning. Subjects assessed are given more attention than ones that are not assessed. When well aligned with content standards, assessments make the intent of the standards explicit and focus attention on content that is deemed important for teachers to teach and for students to learn. But the flip side of that is also true; that is, when there is poor alignment, assessments can distort the intent of the content standards.

Most states make some use of assessment results for school-building accountability. Some emphasize current status. Some use measures of the socioeconomic backgrounds of students attending schools to provide a frame of reference for making comparisons among schools with similar student bodies. The use of socioeconomic measures is controversial, however, because of the implied use of different expectations for students from different backgrounds. Because of the

relationship between these measures and racial and ethnic background of students, the use of socioeconomic measures can have the particularly undesirable result of creating different expectations for White students than for students of color.

A preferable approach for schools that serve students who have low achievement is to place greater emphasis on improvement than on current status. This can be done for schools by comparing the performance of students in a given grade in one year or biennium with that of students in the next year or biennium. Such cross-sectional comparisons are reasonable for schools that serve populations that are fairly stable. Comparisons of the performance of students in a given grade with that of students in the preceding grade the year before can also be used as a way of judging improvement. This can be done for all students in the appropriate grade each year or for only those students with scores in both years. The former approach is known as a quasi-longitudinal analysis, in contrast to the true longitudinal approach with matched student records. Both approaches require tests that have scales that can be compared across grade levels. Both require annual testing in every grade used in the accountability system. Such a requirement is generally associated with the use of either off-the-shelf tests or tests with characteristics similar to off-the-shelf tests, which may suffer from poorer alignment with content standards than assessments that are targeted for just a few selected grades.

The precision of assessment results is less than is commonly assumed by either policymakers or the general public. It is critical that information about the precision of measurement be obtained and provided with reports of assessment results. Given the current emphasis on reporting results for students in terms of whether they meet standards or in terms of a small number of proficiency categories such as below basic, basic, proficient, and advanced, the reports of misclassification probabilities are particularly useful in conveying the level of imprecision in the assessment results. This is true not only at the individual student level but for accountability categories used to classify schools.

A fundamental validity question for any assessment and accountability system is the degree to which results on the assessment generalize to other indicators of achievement. Gains in assessment results for accountability systems have been reported for many states. There are many reasons to expect that the gains reflect many factors in addition to actual improvement in student learning. The narrowing of instruction to the assessments and the widespread use of test preparation

materials can undermine the generalizability of the gains observed. Comparative trends for states from the NAEP provide one of the best ways of evaluating the degree to which inferences from observed increases in scores on state assessments support inferences about improved learning rather than only artificial inflation of scores. Comparisons to other test results such as district-administered norm-referenced tests and to results from college admissions and placement tests such as the ACT, the SAT and Advanced Placement tests are also relevant ways of assessing the degree to which state assessment results generalize.

The preceding review and analysis leads me to offer the following conclusions and recommendations, which I believe will enhance the likelihood that state assessment and accountability systems will contribute to the overarching goal of improving student learning while minimizing some of the potential negative effects that have been discussed.

First, it is important to be clear about the purposes of the assessments and the accountability systems. There is a need to go beyond broad statements that the system is intended to improve student learning. Is the system meant to reinforce content standards in particular subjects because those subjects are judged to be of especially high priority? Is the assessment intended to support deep understanding and ability to solve problems within the subject areas assessed? Are assessment results to be used to assure a given level of achievement for students before they are allowed to move to another grade or to graduate? What type of information will provided to parents? What uses will be made of school-building results? How much emphasis will there be on current performance and how much on improvement?

Once the purposes and intended uses are elaborated, it is critical that the assessments be designed to support the validity of those uses. Key in this regard is alignment of the assessments with the standards. In addition to being aligned, the assessments need to provide good instructional targets for teachers. This implies the need for the introduction of new items and assessment tasks each year, because the intent of the standards that the assessments are supposed to reinforce is clearly broader than a specific set of items and tasks that are administered to students in a given year. Getting by on the cheap, either by using tests that are poorly aligned with the content standards or by the repeated use of the same form of a test year after year, will undermine the value of the assessment since poor alignment distorts instruction, and repeated use of the same form of a test will reduce the validity of results by undermining the generalizability of the results.

To be fair for schools and educators, accountability systems need to place more emphasis on improvement than on current performance. This allows for differences in starting points while maintaining an expectation of improvement for all. High performance standards can also be maintained as a goal for all students.

Reports of results both for individual students and for schools should be accompanied by information about the margin of error in the results. Reporting probabilities that a student or school is misclassified as the consequence of the assessment's measurement error is a good way of conveying the degree of uncertainty that is associated with assessment results.

Because any assessment is fallible, it is unwise to place too much weight on any single test. This implies that when an assessment has high-stakes consequences for a student, the student needs to have multiple opportunities to take the assessment. It also suggests that it is desirable to have multiple ways of assessing the knowledge, understanding, and skills that are the focus of the assessment.

As required by the *Standards* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999), the validity of the uses and interpretations of assessment results needs to be evaluated. Validation also needs to be conducted for the accountability system. The investigation of the degree to which improvements in assessment results generalize to other indicators of achievement is one important aspect of an evaluation of the validity of an accountability system. Another important part of the validation is an evaluation of both the intended positive effects and the more likely unintended negative effects of the system.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bock, R. D., & Wolfe, R. (1996, March 15). *Audit and review of the Tennessee Value-Added Assessment System (TVAAS): Final report.* Nashville, TN: Comptroller of the Treasury.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32,* 385-396.

Burton, E. (2000, April). *A comparison of the generalizability of large-scale performance assessment results for pupil-level and school-level decisions.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

California Department of Education, Policy and Education Division. (2000). *Parent guide to the 1999 similar schools ranks based on the Academic Performance Index.* Sacramento, CA: Author.

Cannel, 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.

Carlson, D. (2000, June). *All students or the ones we taught?* Presentation at the 30th Annual National Conference on Large-Scale Assessment, Council of Chief State School Officers, Snowbird, UT.

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.

Cronbach, L. J. (1963). Course improvements through evaluation. *Teachers College Record, 64,* 672-683.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

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.

Florida Department of Education. (1999). *School accountability report card guide: June 1999.* Available 16 March 2001: http://www.firn.edu/doe/bin00018/guide99.htm

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Hoff, D. J. (2000). As stakes rise, definition of cheating blurs. *Education Week, 19*(41), 1, 14-16.

Hoffman, R. G., & Wise, L. L. (2000). *School classification accuracy final analysis plan for the Commonwealth accountability and testing system.* Alexandria, VA: HumRRO.

Hu, D. (2000). *The relationship of school spending and student academic achievement when achievement is measured by value-added scores.* Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). *What do test scores in Texas tell us?* (DRR-2365-EDU). Santa Monica, CA: RAND.

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

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS)* (MR-1014-EDU). Santa Monica, CA: RAND.

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

Linn, R. L., & Burton, E. (1994). Performance-based assessments: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-8, 15.

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.

Massachusetts Department of Education. (1999). Massachusetts school and district accountability system. Approved September 28, 1999. 603 CMR 2.00, regulations on under-performing schools and school districts, as amended.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14-22.

Meyers, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief, 3*(3). Madison: University of Wisconsin-Madison, National Center for Improving Science Education.

North Carolina Department of Public Instruction. (1996). Setting annual growth standards: "The formula." *Accountability Brief, 1*(1). Raleigh, NC: North Carolina Department of Public Instruction, Division of Accountability Services.

Pennsylvania Department of Education. (n.d.). *Supplemental documentation for 1999: Reading, mathematics and writing assessment reports.* Available 24 April 2001: http://www.pde.psu.edu/pssa/99suppl.pdf

Rogosa, D. (1999a). *Accuracy of individual scores expressed as percentile ranks: Classical test theory calculations* (CSE Tech. Rep. No. 509). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Rogosa, D. (1999b). *Reporting group summary scores in educational assessments: Properties of proportion at or above cut-off (PAC) constructed from instruments with continuous scoring* (Draft Deliverable) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Sanders, W. L., & Horn, S. (n.d.). *An overview of the Tennessee Value-Added Assessment System (TVAAS) with answers to frequently asked questions.* Available 25 April 2001:
http://www.mdk12.org/practices/ensure/tva/tva_1.html

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299-311.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (Research Progress Report). Knoxville: University of Tennessee, Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 137-162) Thousand Oaks, CA: Corwin Press.

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. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450.

Shepard, L., Kupermintz, H., & Linn, R. (2000, February). *Cautions regarding the Sanders value-added assessment system. Response Panel comments.* Presented at the annual conference of the Colorado Staff Development Council, Denver.

Smith, M. S., O'Day, J., & Cohen, D. K. (1990). National curriculum American style: Can it be done? What might it look like? *American Educator, 14,* 10-17, 40-47.

Trimble, C. S. (1994). Ensuring educational accountability. In T. Guskey (Ed.), *High stakes performance assessment: Perspectives on Kentucky's education reform* (pp. 37-54). Thousand Oaks, CA: Corwin Press.