Issues in the Design of Accountability Systems

CSE Technical Report 650

Robert L. Linn
CRESST/University of Colorado at Boulder

April, 2005

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

Project 1.1 Comparative Analyses of Current Assessment and Accountability Systems,
Strand 2: Outcomes of Different Accountability Designs, Robert L. Linn Project Director,
CRESST/University of Colorado at Boulder

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

The findings and opinions expressed do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences or the U.S. Department of Education.

ISSUES IN THE DESIGN OF ACCOUNTABILITY SYSTEMS[1]

Robert L. Linn
CRESST/UNIVERSITY OF COLORADO AT BOULDER

Abstract

The purpose of this report is to identify and clarify design issues that are critical in the creation of an accountability system, and can contribute to improved teaching and student learning. Since student achievement is at the heart of the current accountability movement, a number of issues addressed in this report are concerned with how student achievement is assessed. Also discussed are the many considerations that enter into the choice of measurement and statistical procedures used to estimate school performance. It is concluded that too little attention has been given to the evaluation of current accountability systems and several recommendations are made for improving accountability systems.

Student achievement test results are the coin of the realm in educational accountability systems in the United States. For a number of years, both states and the federal government have relied heavily on tests to judge the quality of schools. The exact characteristics of the accountability systems have evolved over the years and vary a good deal from one state to another as do the state and federal accountability systems. Nonetheless, the systems share, at least implicitly, some underlying beliefs and goals.

Some examples of shared beliefs and goals are:

- The quality of education is not as good as it should be,
- Student learning needs to be improved,
- Student outcomes rather than process or resource measures should be used to judge school quality,

[1] Will also appear as Chapter 4 in: J. L. Herman and E. H. Haertel (Eds.), Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education, vol. 104, Part 1.

- Content standards and associated assessments will make it clear what teachers are expected to teach and students are expected to learn,
- Schools should be held accountable for the learning of all their students,
- Holding schools accountable for student achievement will motivate greater effort on the part of educators and students, and
- Information provided by the accountability system can contribute to improved teaching and student learning.

Despite the shared beliefs and goals there is considerable variation in accountability systems that have been put in place by different states. There are differences in state testing practices such as the grade levels and subjects tested[2] and differences in the stakes that are attached to results (e.g., rewards and sanctions for schools, requiring students to pass the tests for graduation from high school or promotion to the next grade.) There has been a general trend to add high stakes to assessment policies that merely exhort teachers and students to do better (McDonnell, 2004, Chapter 3). When high stakes are attached to results, however, states differ in whether or not there are stakes for students as well as schools.

Some of the differences in assessment and accountability systems are the result of specific legislation, while others are simply the result of administrative decisions and traditions that evolved over time in ways that vary from state-to-state. Some of the differences are also the result of purposeful system design decisions that are intended to meet particular goals.

There are many issues that need to be addressed in the design of an educational accountability system. The purpose of this report is to identify and clarify design issues that are critical in the creation of an accountability system that can contribute to improved teaching and student learning. Since student achievement is at the heart of the current accountability movement, a number of issues that are discussed are concerned with how student achievement is assessed. Other important issues are concerned with the uses that are made of student test results. For example, is the focus on status, or on change, or both? Are results used to make decisions about individual students or only about schools? Issues also arise

[2] As will be discussed below, federal testing requirements under the No Child Left Behind Act of 2002 has introduced some common requirements regarding grades and subjects tested, but states are free to test additional subjects or grades.

in decisions about the reporting of results (e.g., the setting and use of performance standards and rules for disaggregated reports of results).

A Brief History of Accountability Systems

There are two related, but separate tracks, in the development of the schools accountability systems that are now in place in the United States. Federal legislation starting with the Elementary and Secondary Act (ESEA) of 1965 (P.L. 89-10) and continuing with successive reauthorizations of ESEA, the most recent of which is the No Child Left Behind (NCLB) Act of 2002 (P. L. 107-110) provides one track. The second track is comprised of state testing and accountability policies and legislation, and is therefore more variable.

State Accountability Systems

During the 1970s, many states introduced minimum-competency testing requirements for high school graduation and/or grade-to-grade promotion. As the name suggests, the tests generally required only low-level knowledge and skills. Other forms of testing were in most cases left to the districts. In the 1980s, however, many states introduced statewide testing programs. The 1980s was "a decade characterized by deep concern over what was perceived to be poor performance on the part of American students" (Hamilton, 2003, p. 27). Testing was viewed as a useful tool to both monitor and stimulate educational reform efforts. In many states the tests that were used were published standardized tests, however, some states contracted with test publishers to have tests developed that were intended to be a better match to state curriculum guidelines. School level rewards and/or sanctions were introduced by some states.

By the 1990s there was a shift away from the low-level requirements of the minimum competency days. States began to introduce ambitious content standards and assessments that were geared to measure the higher-level understanding and skills called for in the standards. States also started reporting results on assessments in terms of student performance standards, e.g., the percentage of students at the proficient level or above.

Federal Requirements

ESEA was the largest and most enduring of the federal efforts to provide compensatory educational programs for children from low-income families. Title I of ESEA included requirements for testing children receiving Title I services.

Initially, results for Title I students were reported using grade equivalent (GE) scores obtained from the administration of norm referenced tests. The GE scores were found to lack any comparability across test publishers or across subjects for a single publisher. In response to this lack of comparability, the Title I Evaluation and Reporting System (TIERS) was introduced. TIERS required the use of norm-referenced tests and the reporting of results on a new scale, the normal curve equivalent (NCE) scale. NCEs assume a normal distribution of achievement and assign scores from 1 to 99 in a way that makes NCEs of 1, 50 and 99 coincide with the national percentile ranks of 1, 50, and 99, respectively.

By (1994 when the Improving America's Schools Act (IASA) (P.L. 103-382) reauthorizing ESEA was enacted, there was widespread recognition that the reporting of results in terms of NCE scores only for Title I students had major weaknesses. The separate testing of Title I students was often out of sync with the testing of other students and with content standards that a number of states had adopted. IASA required states to adopt content and student performance standards for reporting results on assessments that were supposed to be consistent with the content standards. It also required that states use the same assessments for Title I students as were used by other students in the state. The most recent reauthorization of ESEA is, of course, the No Child Left Behind Act which will be discussed in more detail below.

Assessment of Student Achievement

What is to be assessed is often specified in the most general terms in legislation. NCLB, for example, requires the assessment of students in mathematics and reading or English language arts in Grades 3 through 8 and one high school grade by 2005-2006. By 2007-2008, states will have to assess students in science at least once in each of three grade spans; 3 through 5, 6 through 9, and 10 through 12. NCLB also requires states to set challenging academic content standards and stipulates that the assessments must be aligned with the State's content standards. Legislation in many states has similar requirements. However, such requirements leave a great deal of flexibility in the definition of content standards, the setting of performance standards, and the design of assessments. There is also flexibility with regard to a number of specifics in meeting the requirements of the law such as setting the minimum number of students needed for reporting results for a subgroup to

determine whether specified adequate yearly progress targets have been met by a school or district.

Standards-Based Assessments

Every state except Iowa has established content standards in mathematics and reading or English language arts. The content standards vary substantially in their rigor and specificity, and in some states do not cover all grades (see, for example, Cross, Rebarber, & Torres, 2004; Education Week, 2004). Specificity of content standards is critical if standards are to provide a clear blueprint for the development of assessments. Alignment of assessments with academic content standards is not only called for in legislation, it is essential if assessments are to guide instruction in ways that are intended by the standards.

There is a good deal of evidence that high-stakes assessments influence what teachers teach (see, for example, Hamilton, 2003; Linn, 2003; McDonnell, 2004; Stecher & Hamilton, 2002). Assessments, especially high-stakes assessments, are likely to have more of an impact than content standards on what teachers emphasize. Thus, when content standards and assessments are not aligned, the assessments are likely to distort the intent of the content standards.

Alignment between tests, assessments, and content standards is important for reinforcing the intent of the standards. This important role of alignment led the American Educational Research Association (AERA) to include the following comment in its published position on high-stakes testing.

> "Both the content of the test and the cognitive processes engaged in taking the test should adequately represent the curriculum. High-stakes tests should not be limited to that portion of the relevant curriculum that is easiest to measure. When testing is for school accountability or to influence the curriculum, the test should be aligned with the curriculum as set forth in standards documents representing intended goals of instruction" (AERA, 2000, p. 2).

Evaluating Alignment

Although it is widely agreed that alignment of assessments and content standards is important, there is much less agreement about how alignment should be evaluated. Bhola, Impara, and Budkendahl (2003) reviewed several different approaches that have been used in recent years to evaluate alignment. The approaches differ in complexity. The least complex approach merely matches the

content covered by test items to the content categories of the standards. More complex approaches add consideration of cognitive processes (e.g., Porter, 2002). The evaluation of alignment with respect to cognitive processes is important, because assessments that align well with content standards in terms of coverage of content may do so by overemphasizing factual knowledge and the use of routine procedures rather than conceptual understanding and problem solving skills that are often emphasized in content standards. As was stressed in the AERA statement referenced earlier, assessing only what is easiest to measure is not sufficient. Assessing only factual recall when the content standards stress understanding and problem solving, for example, leads to an assessment target that is poorly aligned with the content standards and is likely to undermine the intent of the standards.

Further complexity may be added to the evaluation of alignment by considering characteristics such as content match, relative emphasis, and depth (similar to cognitive complexity; see, for example, LaMarca, Redfield, Winter, Bailey, & Despriet, 2000; Webb, 1999). Linguistic criteria (e.g., Herman, Webb, & Zuniga, 2002) may also be important to consider along with content and cognitive processes, especially in the assessment of the content knowledge and understanding of English language learners. With any of the approaches, the evaluation of the degree of alignment of assessments and standards is dependent on the level of specificity of the content standards.

Baker (in press) discussed different approaches to alignment in terms of four metaphors: congruence, correspondence (the extent of agreement between assessments and standards), bridge (the connections between assessments and standards), and gravitational pull (generalized processes such as problem solving that cut across content areas). The congruence metaphor corresponds most closely to the way in which alignment is generally evaluated. Complete congruence would require that all instructional goals be clearly specified in the content standards and that each goal be validly measured. Although such complete congruence is unrealistic, the goal in studies of alignment is to evaluate the degree to which it is approximated.

Despite the varied approaches that have been taken, it is clear that the question of alignment has been given much more attention in recent years than it had been in the past. Alignment is clearly an issue that peer reviewers of state responses to NCLB have been attending to and will continue to attend to, thereby assuring continued attention to the topic.

Valid and Reliable Assessments

Legislation mandating assessments and accountability typically include statements that the assessments and/or the accountability judgments must be valid and reliable. As is true of alignment, however, validity and reliability are not all-or-none characteristics. Rather, they are matters of degree. Validity also depends on the specific inferences that are made from assessment results and the uses that are made of the results. For example, an inference that the assessment results reflect the degree to which the content standards have been mastered will depend on the degree to which the assessment is aligned with the standards and therefore has an adequate degree of content validity.

The validity of a conclusion that an improvement in assessment results for a school reflects better achievement of the content standards from one year to the next depends on the degree of alignment, but it also depends on a number of other considerations. For example, it depends on the assumption that the assessment is equally novel for the students taking the assessment in different years. If the same form of a test is used year after year or if there is substantial, say 40% or more, overlap of the items from one year to the next, then student achievement increases may be due to differences in familiarity with the specific test content, and therefore may not generalize to the broader domain of achievement the test is intended to represent. Certainly, the lack of generalization gains shown on a state's assessment of gains on other measures of achievement has led to questions about the validity of the gains (see, for example, Koretz & Barron, 1998; Klein, Hamilton, McCaffrey, & Stecher, 2000; Hamilton, 2003).

Interpretive problems caused by the reuse of tests became evident in the 1980s when most states were reporting upward trends in test scores based on repeated use of the same tests year after year. Indeed, Cannell (1987) reported that almost all states were reporting results above the national average. This "Lake Wobegon effect" (Koretz, 1988) was shown to be the result of inflated scores due to the continual reuse of a single form of a test. When a new test was introduced in a state, there was usually a precipitous drop in the performance on the new test in comparison to that obtained with the old test. Furthermore, the gains found on state tests lacked generalizability to other measures such as the National Assessment of Educational Progress (NAEP) (see, for example, Linn, 2000; Linn, Graue, & Sanders, chapter 7; 1990; Hamilton, 2003).

Most states have moved away from complete reliance on off-the-shelf, norm-referenced tests and have introduced assessments that are designed with the intent of being aligned with the state's content standards prior to the enactment of NCLB. Along with this shift, states are more frequently introducing new items each year. In some cases, however, the core of the assessment still consists of a norm-referenced test that is supplemented with items written specifically for the state in an effort to better align with the content standards. In such cases, there is a need to alternate forms of the core norm-referenced test to avoid excessive year-to-year overlap.

Some items of an assessment have to be repeated from one year to the next in order to equate forms so results in one year can be compared to those of earlier years. This is true whether or not a norm-referenced test is a component of the assessment. A balance is needed between having enough items for dependable year-to-year equating and having so many common items that score gains may be artificially inflated. An overlap of something in the range of a fifth to a quarter of the items will generally be sufficient for equating purposes without substantially undermining the ability to generalize from observed gains to the larger domain of achievement that the assessment is intended to measure.

As is true of validity, reliability also needs to be evaluated in relationship to the uses and interpretations of assessment results. For school accountability uses, the reliability of individual scores, while of concern for appropriate interpretations of student scores, is not the main issue. Rather, the primary concern is with the consistency of results used to make judgments about schools. Thus, when NCLB requires that adequate yearly progress (AYP) be defined in a manner that is statistically reliable (NCLB, 2002, § 1111(b)(2)(C)(ii-iii)), it should be interpreted to mean that schools should be consistently classified as meeting or failing to meet AYP targets. Consistency of school classification in terms of AYP status depends on the number of students in the schools. Small schools will be less reliably classified than large schools (Hill & DePascale, 2003; Linn & Haug, 2002). Since schools can fail to meet the AYP target because any single subgroup for which disaggregated results must be reported fails to meet the target, it is also the case that schools with many subgroups of students of sufficient size to be reported will also be classified less reliably than schools with a more homogeneous student body (Hill & DePascale, 2003; Kane, Staiger, & Geppert, 2002).

Accountability Uses of Student Assessment Results

Student assessment results are used in several different ways by systems that have been put in place by states as part of their state-mandated accountability or in response to NCLB. Student assessment results for particular grades and subjects aggregated at the school level may be compared to a performance target. The emphasis may be on current status or on improvement. Improvement may be judged by comparing the performance of successive cohorts of students (e.g., fourth grade students in 2004 compared to fourth grade students in 2003), or it may be judged by tracking student performance longitudinally (e.g., the comparison of the performance of the same students as fifth graders in 2004 to their performance as fourth graders in 2003). Some systems focus on each content area assessed while other systems rely on composite results across content areas. State systems may focus only on results for the school as a whole or, as is required by NCLB, may also set requirements for subgroups of students defined by race/ethnicity, socio-economic status, disability status, and English language proficiency.

Status Measures

Reporting results for a school in terms of current status (e.g., the percentage of students at the proficient level, or an index score determined by the percentage of students in each of several performance categories) is part of almost all state accountability systems and is required by NCLB. Reliance on current status measures only, however, raises a number of issues of fairness to schools. Current status is

> …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).

On the other hand, current status indicators have the perceived advantage that they hold all students and schools to the same standards of performance. They, obviously, also provide information about where students and schools stand at a given point in time, and when current status is compared to performance targets, how far there is to go.

NCLB holds schools accountable for meeting AYP targets, which are called annual measurable objectives (AMOs). Despite the fact that the "P" in AYP stands

for progress, in any given year, it is only current status compared to the AMOs that determines whether or not a school has met AYP requirements.[3] Because current status measures place schools serving poor and/or initially low-achieving students at a disadvantage in comparison to schools serving more affluent and/or initially high-achieving students, most state accountability systems consider improvement of some form in addition to current status.

Improvement Measures

Successive cohorts. The most common approach to measuring improvement is to compare the performance of successive cohorts of students in the same school. The successive cohort approach to measuring improvement has the advantage of simplicity. It does not require the tracking of individual students from one year to the next. This approach has its limitations, however. It does not account for changes in the student characteristics from one year to the next. The performance of students who transfer into a school shortly before the date that assessments are administered cannot reasonably be attributed to instruction provided by the school, but they will help determine school gains or losses unless they are explicitly excluded in the calculation of year-to-year change. Change or difference scores are also less reliable than the individual scores that are used to compute the difference. Consequently, schools that show minimal gains or even losses in one change cycle will frequently have large gains in the next change cycle. Conversely, schools with outstanding gains in one change cycle will often show only small gains, or may even have losses, in the next change cycle (see, for example, Haney, 2002; Linn & Haug, 2002).

Longitudinal measures of change. Tracking individual students from year-to-year allows each student to serve as his or her own control. Longitudinal measures of student growth are appealing to schools and teachers, in part, because they generally only hold the school accountable for students that are enrolled in the school for at least a year. On the other hand, exclusion of mobile students is viewed as a disadvantage by those that want all students included in the accountability system.

---

[3] NCLB does have a safe harbor provision that allows a school that would not otherwise meet AYP to still be classified as meeting AYP if (1) the percentage of students who score below the proficient level is decreased by a least 10% from the year before, and (2) there is improvement for the subgroup on other indicators. In practice, however, few schools are "saved" by the safe harbor provision because the bar is set so high.

Student growth may be monitored over a pair of years (e.g., fourth to fifth grade) or over multiple years (e.g., third, fourth, and fifth grades). Simple difference scores or complex statistical analyses may be used to estimate growth. In either case, the scores at one grade level need to be comparable to those in other grade levels included in the analysis for the use of gain scores.

Because a test appropriate for, say, sixth grade students contains more challenging content than one appropriate for fourth grade students, simple scores such as the number of correct responses are not comparable for tests used at different grade levels. Typically, scores at different grade levels are made comparable by constructing a vertical scale, i.e., a scale that places scores on tests used at different grade levels on a common numerical scale. A vertical scale is usually created by including a common set of anchor items in adjacent grades. When item response theory models are used, the unique items at each grade level are put on the common vertical scale by fixing the anchor item parameter estimates. Similarly, if classical scaling procedures are used, the statistics for the anchor item set are fixed so that item sets unique to a grade can be placed on the vertical scale.

Mathematics assessments at Grades 3 and 8 obviously have dramatically different content and require quite different skills. It is implicitly assumed, however, that a vertical scale that spans Grades 3 through 8 measures a common construct. In fact, however, the construct changes over grade levels. For example, the factual and procedural knowledge required to add and subtract and recognize geometric figures at the third-grade level may differ substantially from the conceptual understanding and pre-algebra problem solving skills required in the eighth grade. Such changes in the complexity of the construct that is measured make difference scores across a wide span of grade levels difficult to interpret (see, for example, Reckase, 2004). On the other hand, difference scores on a vertical scale for adjacent grades or a small span of, say, three grade levels are easier to interpret because of the greater similarity of test content and cognitive processes required of students by assessments designed for use in adjacent grades.

In addition to the need to attend to shifts in the construct measured across grade levels, it is important to understand the statistical properties of vertical scales. It is particularly important to attend to the relative size of the variance in vertical scale scores from one grade to the next. Vertical scales have on occasion had variances that decreased with grade level, had variances that were approximately equal, or had variances that increased with grade level. Gain scores, e.g., the

difference between the vertical scale scores obtained by sixth-grade students in 2004 and the scores they obtained as fourth-grade students in 2002, clearly would have different interpretations depending on whether variability decreases, increases, or remains constant across grade levels.

Issues raised by changes in the construct measured and changes in the variability of scores over grades are most evident in the interpretation of simple difference scores. These issues, however, need careful consideration regardless of whether simple difference scores or more sophisticated statistical models are used to estimate growth in student achievement.

Analyses of longitudinal data. There has been considerable interest in recent years in using complex statistical models to analyze longitudinal student achievement test data collected over several years to estimate school and teacher contributions to student gains in achievement. The most prominent example of this type of analysis is the Tennessee Value Added Assessment System (TVAAS) (Sanders & Horn, 1998). The "value added" terminology is used to convey the notion that the estimated teacher and/or school effects represent the contributions made by either a teacher or a school to student gains on achievement tests beyond that expected based on student performance for the past three years. The analytic models employed by Sanders have attracted a great deal of attention across the country. The widespread interest is due, in part, to findings reported by Sanders and his colleagues that purport to show that teacher contributions to student learning are quite large compared to other factors (Wright, Horn, & Sanders, 1997) and that teacher effects accumulate over time (Sanders & Rivers, 1996).

The interest in value-added models has mushroomed since the approach was introduced on a large scale in Tennessee a decade ago. As noted by Wainer (2004) in his introduction to the special issue of the *Journal of Educational and Behavioral Statistics* on value-added models, several states in addition to Tennessee already use some form of longitudinal measurement and several more are either exploring or have already mandated the use of value-added models with longitudinal data.

> Enthusiasm for this approach stems in large part from the belief that it can remove the effects of factors not under the control of the school, such as prior performance and socioeconomic status, and thereby provides a more accurate indicator of school or teacher effectiveness than is possible when these factors

are not controlled (McCaffrey, Lockwood, Koretz, Lewis, & Hamilton, 2004, p. 68).

For a more extended discussion see McCaffrey, Lockwood, Koretz, and Hamilton (2003).

Statements regarding the "effectiveness" of teachers or schools based on results of value-added analyses are basically causal claims. Rubin, Stuart, and Zanutto (2004) have argued, however, that value-added analyses "should not be seen as estimating causal effects of teachers or schools, but rather as descriptive measures" (p. 113). In a similar vein, Raudenbush (2004) cautions that although value-added analyses "when combined with other information have potential to stimulate discussions about how to improve practice they should not be taken as direct evidence of the effects of instructional practice" (p. 128). Ballou (2004) provides a different perspective in his rejoinder to the arguments made by Raudenbush and by Rubin et al arguing that the estimates done properly will not necessarily be biased and noting that all models are subject to criticism.

Reporting and Interpreting Results

A defensible and useful educational accountability system based on student achievement requires a valid and reliable measurement of the intended content domain. It also requires well-defined, systematic procedures for analyzing student achievement data to make inferences about the performance of schools. As has been discussed in previous sections of this report, there are many considerations that enter into the choice of measurement procedures and statistical procedures used to estimate overall school performance. In addition to the measurement and analytic considerations, the validity and utility of the accountability results depend on the ways in which the results are reported and interpreted.

Performance Standards

Approaches to reporting achievement test results that were used in the past, such as national percentile ranks, scale, grade-equivalent, or normal-curve equivalent scores, fell in to disfavor in the last decade. Such reporting has generally been replaced by standards-based score reports that present results in terms of percentages of students in performance categories, e.g., below basic, basic, proficient, and advanced. Alternatively, or in addition, percentages of students who score above a given level (e.g., proficient or above) may be reported. The switch to

performance standards as a reporting mechanism was motivated by the desire to go beyond normative statements about performance to answer the question: how good is good enough?

The interest in performance standards developed fairly rapidly at the national level and in a number of states. The Goals 2000: Educate America Act of 1994 (P.L. 103-227) and the Improving America's Schools Act of 1994 (P.L. 103-382) called for performance standards which were intended to specify the level of achievement that should be considered good enough—typically referred to as proficient. Performance standards, referred to as achievement levels, were set on NAEP by the National Assessment Governing Board. A number of states also set performance standards, often modeled after NAEP, on their state assessments.

These performance standards had several common properties. They were absolute rather than normative. They were set in a context that called for ambitious, "world-class" standards, and they divided the range of student achievement on an assessment into a relatively small number (typically 4) of categories. Finally, they were expected to apply to all, or nearly all, students.

NCLB requires states to set "challenging student academic achievement standards" (P. L. 107-110, Section 1111(b)(1)(A)) for their state assessments in mathematics and reading/English language arts. States must set at least three performance standards for their assessments to comply with NCLB requirements. Two of the standards (proficient and advanced) are supposed to correspond to high levels of achievement and the third standard (basic) is intended to provide a means of monitoring progress toward proficient achievement. Most states had already set performance standards for their reading and mathematics assessments before NCLB was signed into law in January, 2002, albeit not necessarily at all the grade levels that must be assessed by 2005-06. Since that time the other states have set performance standards for their assessments and, in a few instances, states that had already set standards have revised or renamed their standards and taken steps to set them at grades where they did not have assessments that will be required under NCLB. It is notable that the state standards vary greatly in stringency and that standards set by states after NCLB became law tend to be less stringent than standards set before 2002, presumably because of the sanctions imposed by NCLB for schools where the percentage of students meeting standards falls below established annual targets.

Status and Improvement Targets

As was noted previously, the context in which performance standards were set in the 1990s on NAEP and on state assessments was one in which there was considerable discussion of high, world-class performance expectations. Not surprisingly, the standards tended to be set at quite high levels (see, for example, Shepard, Glaser, Linn, & Bohrnstedt, 1993). It is also worth noting that in the NAEP performance standards and standards on many state assessments served only hortatory purposes, and had no real consequences for students or schools prior to the enactment of NCLB.

NCLB and state expectations. More recent state uses of performance standards have had consequences for schools and/or students and there certainly are consequences for schools and districts as under NCLB. NCLB specifies state procedures used to set AMOs based on the percentage of students performing at the proficient level, or above, that are to be applied to determine if schools, districts, and states make AYP. The AYP targets must be set such that all students will be at the proficient level or above by 2014. Sanctions are imposed on schools not meeting their AYP targets two years in a row and the sanctions become increasingly severe for schools not meeting targets for a third, fourth, and fifth year in a row.

If the NCLB requirement that states set standards at challenging levels is followed and the proficient standard is set at a high level, then the mandate that all students perform at the proficient level or higher by 2014 will be completely unrealistic. Using trends on NAEP as a benchmark, Linn (2003, 2004) has shown that the rate of improvement in student achievement would have to be many times faster for the next decade than it has been for any comparable period of time in the last 40 years. Unless the AYP targets that have been set assuming the achievement of the 100% proficient or above goal by 2014 are changed, nearly all schools will fail to meet their AYP targets within the next few years.

Ambitious but realistic expectations. Ambitious expectations are desirable to encourage concentrated effort on the part of educators and students. However, in order for the expectations to be met, educators and students must have the capacity to meet the targets that are set. Effort alone is insufficient if teachers lack the knowledge they need to meet expectations. Exhortation may increase motivation, but is no substitute for capacity building. Even added resources, capacity building, and increased effort, however, cannot lead to the accomplishment of goals that are

set at unrealistically high levels. Thus, it is critical that expectations be set at ambitious but realistic levels.

One way to ensure that performance goals are both ambitious and realistic is to base the goals on the accomplishments of schools displaying the most rapid gains in achievement over a period of several years. If the rates of improvement that the top, say, 10% of schools have achieved over the past five years were set as expectations for all schools, for example, the goals would certainly be ambitious for the vast majority of schools, but they would also be more realistic than expectations such as those set by NCLB without regard to what has been achieved in the past by exemplary schools. Note that by focusing on sustained improvement rather than status, schools that are successful in teaching students living in poverty, not just schools serving upper middle class students who perform at high levels in a given year, could be the ones that set the growth targets for other schools.

Subgroup performance. One of the important features of the NCLB accountability system is the requirement to report results separately for students with disabilities, economically disadvantaged students, students with limited English proficiency, and by race/ethnicity. Such disaggregated reporting of results was already required by the accountability systems in a few states, but with the enactment of NCLB, is now required in all states. Reporting performance for the subgroups specified by NCLB is important for ensuring that attention be given to the achievement of groups of students that have too often been ignored in the past. It is also critical for monitoring the degree to which the achievement has been reduced.

Although it is highly desirable that disaggregated reporting of student achievement results required under NCLB be continued, there is a problem caused by this requirement that needs to be addressed. Since schools can fail to meet AYP in many different ways but can meet it in only one way, it turns out that schools with a sufficient number of students in each of several targeted groups to be reported are less likely to meet AYP targets than schools of the same size and similar performance but with a homogeneous student body (e.g., nearly all students who belong to one racial/ethnic group).

The most straight-forward way of avoiding the over identification of schools with multiple subgroups as failing to make AYP would be to modify the safe harbor provision of NCLB. This provision allows schools where a subgroup of students falls

short of the AYP target to still meet AYP if (1) the percentage of students in that subgroup who score below the proficient level is decreased by 10% from the year before, and (2) there is improvement for that subgroup on other indicators.

Because the 10% reduction in students scoring below the proficient level is a very high bar, very few schools that would not otherwise make AYP do so because of the safe harbor provision. Only a tiny fraction of schools actually meet AYP through the safe harbor provision because it is so extreme. Changing the safe-harbor provision from a 10% reduction in below proficient to, say, a 3 or 4% reduction would go a long way toward solving the problems caused by the multiple hurdles created by subgroup reporting while assuring improvement in performance of all subgroups.

Value-Added Reporting

Value-added models yield measures of individual student gains in achievement, and when those gains are linked to the students' schools and teachers, those gains are converted into estimates of school effects and teacher effects. Gains may be expressed in a variety of metrics, including scale scores or comparisons to national norms.

Schools. TVAAS reports estimates of school gains in comparison to state and national gains. Mean gains are reported for a school by grade level and subject area (reading, mathematics, language, science and social studies). Those mean gains are compared to gains statewide and gains from one grade to the next in national norms. Cumulative percents of national norm gains are also reported. A cumulative percent of a national norm of 110, for example, indicates that the school is gaining at a rate that is faster than the national norm. Although the analyses that yield the estimates of gain for a school are complex, the descriptive statement that the gain exceeds the national norms gain is straight-forward and not subject to much, if any, controversy.

It is the interpretation of the gains as an indication of school effectiveness that is controversial. As was discussed above, such an interpretation is fundamentally a causal claim, and strong and questionable assumptions must be satisfied to make causal inferences about school or program effectiveness from the results of value-added analyses. Detailed explanations for this conclusion about the inability of value-added analyses to support causal inferences are provided by Raudenbush (2004) and by Rubin, Stuart, and Zanutto (2004), but see also, Ballou (2004) who

noted that the specification of any model may be questioned. He also suggested that, by making allowances for students' starting level, value-added models are better than the approach of NCLB which makes no such allowances.

Teachers. Not all applications of value-added models include estimates of teacher "effects." Indeed, some state mandates, e.g., Ohio, are by design limited to the school with no intention of estimating gains associated with individual teachers. Individual estimates of "teacher effects" are produced by TVASS, but they are intended for use by the teacher and are not included in public reports. The estimated "[t]eacher effects are deviations from the district average" within a given year (Ballou, Sanders, & Wright, 2004, p. 40). As in the case of estimated "school effects," causal interpretations of "teacher effects" can be justified only if strong, unverifiable assumptions are met.

Students. Measures of student gains in achievement obtained from value-added analyses are far less controversial than the measures obtained for either teachers or schools. This is so, because the gains are treated as descriptive information without any causal attributions. The goal is usually to determine if a student has made a year's worth of growth in a year or to indicate that the rate of growth is sufficient to achieve some target such as achievement at a proficient or advanced level by some specified grade.

Summary and Conclusion

External assessments of student achievement have been used for purposes of school accountability for more than half a century. The specific ways that assessments have been used have expanded and changed over the years. As McDonnell (2004) has noted, "assessment has become a critical accountability tool" (p. 9). There are several reasons for the appeal of assessment-based accountability (see, for example, Linn, 2000; McDonnell, 2004).

Assessments, unlike changes in instructional practices, can be mandated at the state level and there is reasonable assurance that they will be implemented in ways that are generally consistent with adopted policies. Compared to other alternative policies intended to improve student achievement, such as reducing class size or adding tutors, assessments are also relatively inexpensive. Assessments can be implemented in a relatively short period of time and the results are visible. In addition to their hortatory value, it is relatively easy to increase the stakes associated with assessment results by adding rewards and sanctions.

Despite the clear appeal of assessment-based accountability and the widespread use of this approach, the development of assessments that are aligned with content standards and for which there is solid evidence of validity and reliability is a challenging endeavor. Alignment of an assessment with the content standards that it is intended to measure is critical if the assessment is to buttress rather than undermine the standards. Too little attention has been given to the evaluation of the alignment of assessments and standards.

Evaluations of alignment can provide support to validity claims regarding content, but other types of evidence are needed to support the myriad interpretations of assessments within an accountability system. The conclusion that improved performance on an assessment implies better learning of the content domain, for example, requires evidence that the gains generalize to other indicators of achievement in the content domain.

Moving from assessment results for individual students that may have good validity as measures of student achievement in a content area to aggregate results that draw inferences about the quality of schools, poses additional validity questions. Although the finding that a large percentage of students score at the proficient level or above may provide an accurate indication of the achievement of students within a school, it cannot be simply converted to a conclusion that the high achievement is the result of good instruction or that it means the school is of high quality. The high achievement may reflect prior achievement and other characteristics of the student body, as much so or more than it does the quality of instruction in a given year.

Similarly, gains in achievement for successive cohorts of students in a school may be due to improved instruction, but they may also be due to a variety of other factors such as changes in the mix of students, or a narrow focus on the specific assessment. Longitudinal student data can rule out the possibility that gains are due to changes in the mix of students from one year to the next. The use of sophisticated value-added analyses can eliminate some of the possible interpretations of gains. However, as was discussed above, causal claims that the gains are due to school or teacher effects can still be challenged.

Performance standards were introduced as a means of reporting results, in part, because it was thought they would make results easier to interpret and, in part, to set expectations for acceptable levels of achievement. Because of the huge

variability in the stringency of performance standards from state to state, it is not clear that standards-based reporting has made results more readily interpretable.

The high level of performance standards that is encouraged by NCLB and that has been put in place on assessments in a number of states, together with the NCLB mandate that all students be at the proficient level or above by 2014, has resulted in expectations that are quite unrealistic. Something will have to give. Either the expectations will be made more realistic or nearly all schools will fall short of AYP targets within the next few years.

Several characteristics of accountability systems are important if the system is going to have its intended positive effects on teaching and student learning. Assessments need to be aligned with content standards, which in turn need to provide clear indications of the content to be taught, and to the cognitive processes that students are expected to use in demonstrating understanding and solving problems. Both status and improvement should be considered in the accountability system. Attention needs to be given to the performance of student subgroups who have lagged behind their better-off peers in the past. The jury is still out on what combination of rewards and sanctions is most effective. It is clear, however, that teachers in schools that are not performing well need assistance and sustained professional development so that they can better do the job that they almost all would like to be doing: more effectively facilitate the learning of all their students.

References

American Educational Research Association. (2000). *AERA Position statement concerning High-stakes testing in preK-12 education*. (http://www.aera.net/about/ploicy/stakes.htm).

Baker, E. L. (in press). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ballow, D. (2004). Rejoinder. *Journal of Educational and Behavioral Statistics*, *29* (1), 131-134.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29* (1), 37-65.

Bhola, D. S., Impara, J. C., & Buckendahl, W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and practice*. *22*, No. 3, 21-29.

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

Cross, R. W., Reparber, T, & Torres, J. (Eds.). (2004). *Grading the systems: The guide to state standards, tests, and accountability policies*. Washington, DC: The Fordham Foundation and Accountability Works (http://www.edexcellence.net/foundation/publication.cfm?id=328).

Education Week (2004). Quality Counts 2004. Volume 33, No. 17, January 8.

Elementary and Secondary Education Act of 1965, Public Law 89.10.

Goals 2000: Educate America Act of 1994, Public Law 103-227.

Hamilton, L. (2003). Assessment as a policy tool. In R. E. Floden (Ed.), *Review of Research in Education*, *27*, pp. 25-68.

Haney, W. (2002, May 6). Lake Woebeguarenteed: misuse of test scores in Massachusetts, Part I. *Education Policy Analysis Archives*, *10* (24).

Herman, J. L., Webb, N., & Zuniga, S. (2002). Alignment and college admissions: *The match of expectations, assessments, and educator perspectives*. Paper presented at the

annual meeting of the American Educational Research Association, New Orleans, LA.

Hill, R. K. & DePascale, C. A. (2003, April). *Adequate yearly progress under NCLB: Reliability considerations*. Paper presented at the 2003 Annual meeting of the National Council on Measurement in Education, Chicago. Retrieved from http://www.nciea.org/.

Improving America's Schools Act of 1994, Public Law 103-382.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). Randomly accountable. Retrieved from http://www.educationnext.org/20021/56.html.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* 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. M. & Barron, S. I. (1998). *The validity of gains in score on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

La Marca, P., Redfield, D., Winter, P., Bailey, A., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of Chief State School Officers.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29* (2), 4-14

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researche*r, *32*(7), 3-13.

Linn, R. L. (2004). *Rethinking the No Child Left Behind Accountability System*. Washington, DC: Center on Education Policy.

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*, no. 3, 5-14.

Linn, R. L. & Haug, C. (2002). Stability of school building scores and gains. *Educational Evaluation and Policy Analysis*, *24*(1), 27-36.

McCaffrey, E. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McCaffrey, E. F., Lockwood, J. R., Koretz, D. M., Lewis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29* (1), 67-101.

McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

Meyers, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief*, 3, No. 3. Madison, WI: National Center for Improving Science Education, University of Wisconsin-Madison.

No Child Left Behind Act of 2002. Public Law 107-110.

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

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29* (1), 121-129.

Reckase, M. D. (2004). The real world is more complicated than we would like it to be. *Journal of Educational and Behavioral Statistics*, *29* (1), 117-120.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment. *Journal of Educational and Behavioral Statistics*, *29* (1), 103-116.

Sanders, W. & Horn, S. (1998). Research findings from the Tennessee value added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

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

Shepard, L., Glaser, R., Linn, R. & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.

Stecher, B. M. & Hamilton, L. S. (2002). Putting theory to the test: Systems of "educational accountability should be held accountable". *Rand Review*, *26*(1), 16-23.

Wainer, H. (2004). Introduction to a special issue of the *Journal of Educational and Behavioral Statistics on value-added assessment*. *Journal of Educational and Behavioral Statistics*, *29* (1), 1-3.

Webb, N. L. (1999). Research Monograph No. 18: *Alignment of science and mathematics standards and assessments in four states*. Madison WI: National Institute for Science Education.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation*, *11*, 57-67.