Educational Accountability Systems

CSE Technical Report 687

Robert L. Linn
CRESST/University of Colorado at Boulder

June 2006

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
Robert 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.

EDUCATIONAL ACCOUNTABILITY SYSTEMS

Robert L. Linn
CRESST/University of Colorado at Boulder

Abstract

Test-based educational accountability systems have considerable appeal to politicians, policymakers, and the general public. Such systems have been widely used by states for more than a decade and with the enactment of the No Child Left Behind Act of 2001 all states must now implement an accountability system that uses results from assessments in mathematics and English/language arts and is administered each year in grades 3 through 8 plus one high school grade. A wide variety of test-based accountability systems of states and as required by NCLB are described and their strengths and weaknesses are evaluated. It is argued that the sanctions for schools that are part of the accountability system require causal inferences about school effectiveness. It is concluded, however, that basing causal inferences about school quality on the results that can be obtained from the existing school accountability systems is not scientifically defensible. It would be better to view accountability results as a source of descriptive information about schools and the basis of hypotheses that can be evaluated by gathering addition information about instructional staff and practice.

Educational Accountability Systems

When President Bush made accountability the centerpiece of the education agenda that he outlined at his January 23, 2001 press conference on education (The White House, 2001) he reinforced what was already a central theme of state polices aimed at improving education. Many of the accountability features of President Bush's education agenda were incorporated into the "No Child Left Behind Act of 2001" (NCLB) that President Bush signed into law in January, 2002.

NCLB amends the Elementary and Secondary Education Act (ESEA) of 1965 and provides significant financial support to schools and districts serving students from low-income families. The law was enacted with strong bipartisan support in both houses of Congress. NCLB's testing and accountability provisions are consistent with the broad outline provided in President Bush's January 23, 2001

press conference, including, for example, the requirement that students be tested every year in reading and mathematics in grades 3 through 8.

The NCLB accountability system as well as accountability systems that were put in place by a substantial number of states in the 1990s reflect the beliefs of politicians, policymakers, and the business community that educational achievement is inadequate and unequal and that educational reforms are needed (McDonnell, 2005). But, as Rouse (2005) has noted, policymakers also believe "that traditional forms of school improvement, such as class-size reduction and professional development, are expensive and ineffective" (p. 275; see also McDonnell, 2004, pp. 9-10). And, they believe that a lack of school accountability has contributed to the poor performance of schools. Hence, they have turned to test-based school accountability as a major component of educational reform, reasoning that sanctions and rewards to schools will prod teachers and school administrators to be more effective (McDonnell, 2005; Rouse, 2005). President Bush's call for frequent testing of students in reading and mathematics is consistent with this set of beliefs, but that position did not originate with President Bush or with the enactment of NCLB.

Rationales for Testing

There is nothing novel about the idea of using student achievement test results as a major component to an educational accountability system. As recently documented by Haertel and Herman (2005), a large number of accountability testing programs have been introduced over the past century. The uses of test results and rationales for testing have varied, but there have been several common and recurring themes. Among other things, tests have been expected to:

- help clarify expectations for teaching and learning;
- monitor educational progress of schools and students;
- monitor the progress of demographic subgroups of students and the gaps in achievement of those subgroups;
- encourage the closing of the gaps in the performance among racial/ethnic subgroups and between economically disadvantaged students and their more affluent peers;

- motivate greater effort on the part of students, teachers, and school administrators;
- contribute to the evaluation of educational programs and schools;
- identify schools and programs that need to be improved; and
- provide a basis for the distribution of rewards and sanctions to schools and students.

Tests are used as policy tools to hold teachers and school administrators accountable for student learning and as levers to change instruction in the classroom. As McDonnell (2004) has noted, "elected officials and policymakers have few tools at their disposal [to change classroom instruction]. Standardized tests are one of the most effective levers they have for influencing what happens in local schools and classrooms" (p. 9).

Assessment of Student Achievement

Achievement tests have been at the core of outcome-based accountability systems of school districts, states, and the federal government for a number of years. The nature of the tests and the uses that are made of test results have not been uniform, however. Achievement tests have undergone several changes over the years. There have been periods when multiple-choice, norm-referenced achievement tests dominated the educational testing and accountability scenes. Basic skills were stressed during the minimum-competency testing period of the late 1970s and early 1980s (Hamilton, 2003; Linn, 2000). Worries that an exclusive emphasis on basic skills could make the minimum become the maximum led to calls for giving greater attention to more complex comprehension and problem solving skills. The publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983) is particularly notable in this regard.

A Nation at Risk marked the beginning of a turning point in educational testing and accountability. It planted the seeds for the standards and performance assessment movements (Haertel & Herman, 2005). The use of performance assessments, requiring students to write extended responses, solve real-world mathematics problems and defend their solutions, was introduced in the late 1980s and early 1990s. Performance assessments stressed depth of understanding and complex problem solving skills. They were expected to have a major impact on education by providing models for teaching and learning.

Although there was a widely held belief that the use of performance assessments would have instructional advantages over more traditional multiple-choice and short-answer test items, the heavy use of performance assessments faded fairly rapidly due to technical quality and cost issues. The standards movement spawned by *A Nation at Risk* and originally closely associated with performance-based assessments, however, has continued and, if anything, played an increasingly central role testing and accountability.

The standards movement of the 1990s was championed at the federal level by the Clinton administration's education initiative articulated in the *Goals 2000: Educate America Act*. Content standards, student performance standards, and standards-based assessments were key ideas in this initiative. The standards-based approach to assessment and accountability was reinforced by the requirements for Title I evaluations mandated in the 1994 re-authorization of ESSA by the *Improving America's Schools Act of 1994* (the predecessor to NCLB). Because of this earlier push for standards, a majority of the states had adopted or were working toward a standards-based approach prior to the enactment of NCLB. Although only 9 states were reported to have standards-based tests in both reading/English language arts and mathematics at grades 3 through 8 in 2002 (Olson, 2002), most states had adopted content standards in those subjects and had tests that were arguably aligned with those standards at some grade levels.

Content and Performance Standards

By the 2004-2005 school year, every state except Iowa had adopted content standards in core subjects (Education Week, 2005, p. 86). With few exceptions the core subjects for which content standards had been adopted included mathematics and reading or English language arts as required by NCLB. Forty-four states reported that they had developed tests that were customized to be aligned with their content standards and an additional 5 states reported that they had augmented or hybrid tests that were aligned with their content standards (Education Week, 2005, p. 86). The adequacy of the alignment can, of course, be questioned. Evaluation of the alignment of tests with content standards will be discussed in the following section.

NCLB requires states to adopt "challenging academic content standards." Those standards are supposed to "specify what children are expected to know and be able to do; contain coherent and rigorous content; [and] encourage the teaching of

advanced skills" (NCLB, 2002, part A, subpart 1, Sec. 1111, a (D). The required academic content standards are expected to provide the basis for developing challenging assessments that are well aligned with the content standards. They also are expected to provide the framework for specifying "challenging student academic achievement standards" that are set at a minimum of three levels, usually called "advanced," "proficient," and "basic," that divide the students into four categories—advanced, proficient, basic, and below basic. NCLB requires that the same set of student academic achievement standards be applied to all students.

Student academic achievement standards are more commonly referred to as performance standards. Although performance standards are dependent upon academic content standards, they are distinct. The content standards specify the "what" while performance standards specify the level of achievement that is expected of students. According to NCLB, students must demonstrate high levels of achievement to meet the proficient or advanced performance standard. Yet, all students are expected to perform at the proficient level or above by 2013-2014.

Performance standards became popular as a way of reporting prior to the enactment of NCLB (Hamilton, 2003). They were identified as an alternative way of reporting test results in *Goals 2000*. Reporting percentages of students meeting or exceeding performance standards was seen as preferable to reporting means or normative information because the standards were supposed to specify a level of performance that was judged to be good enough or exemplary without referring to the performance of other students.

Performance standards have at least five critical characteristics. First, they are intended to be absolute rather normative. That is, they are supposed to establish a fixed criterion of performance. Second, they are supposed to be challenging, that is, to be set to correspond to high levels of achievement. Third, a relatively small number of levels (e.g., advanced, proficient, and basic) are typically identified. Fourth, they are expected to apply to all, or essentially all, students rather than a selected subset such as college-bound students seeking advanced placement. Finally, they depend on judgments made by standard setters who review statements about expected levels of achievement and translate those statements into cut scores on assessments (Linn, 2005b).

It is easy enough to talk about basic, proficient, and advanced levels of performance in terms of the knowledge and skills specified in the content standards.

Performance levels must be mapped onto scores on an assessment to have meaning, however, in terms of actual student achievement. Mapping performance levels onto an assessment has a great deal of uncertainty associated with it. Where the standards are set depends on characteristics of the judges who set the standards, and on the method used to set the standards (Linn, 2003). The stringency of the standards also depends on the context in which the standards are set. States that set standards before NCLB when there were no stakes attached to falling below the proficient cut, for example, often set the standards at quite ambitious levels. The judges setting standards in states since NCLB became law, on the other hand, were aware of the consequences of students failing to reach the proficient level and have generally set more lenient standards.

The variability in the stringency of the state standards defining proficient performance is so great that the concept of proficient achievement lacks meaning (Linn, 2003). Although one would not expect state NAEP results to perfectly track results on the assessments used by different states, NAEP does provide a reasonable benchmark for comparing the standards set by different states. Olsen (2005) reported the percentages of students who were at the proficient level or above on the 2005 state grade 8 mathematics assessments for the 33 states that administered mathematics assessments at that grade and for which results were available in time for her report. She also listed the percentage of students who scored at the proficient level or above on the 2005 grade 8 NAEP for those states. A plot of the differences in the percentages of students who were proficient or above on the state assessments minus the corresponding percentages on the NAEP assessment is displayed in Figure 1. The mean difference in percentage was 33%, i.e., 33% more students scored at the proficient level or above according to the state assessment than did so according to NAEP. The difference ranged from a low of -10% for Missouri to a high of 66% for Tennessee, suggesting that the Missouri proficient standard is somewhat more stringent than the NAEP standard while the Tennessee standard is much more lenient than NAEP.

A the relationship between the percentages of students who were proficient or above in 2005 according to the state grade 8 mathematics assessments and the corresponding percentages according NAEP is illustrated by the scatterplot shown in Figure 2. As can be seen in Figure 2 the relationship between the percentage of students who are proficient or above according state assessments and NAEP is relatively weak ($r = .34$). It is also apparent that the variability in percentages is

much greater on the state assessments (standard deviation = 18.4) than on NAEP (standard deviation = 6.7).

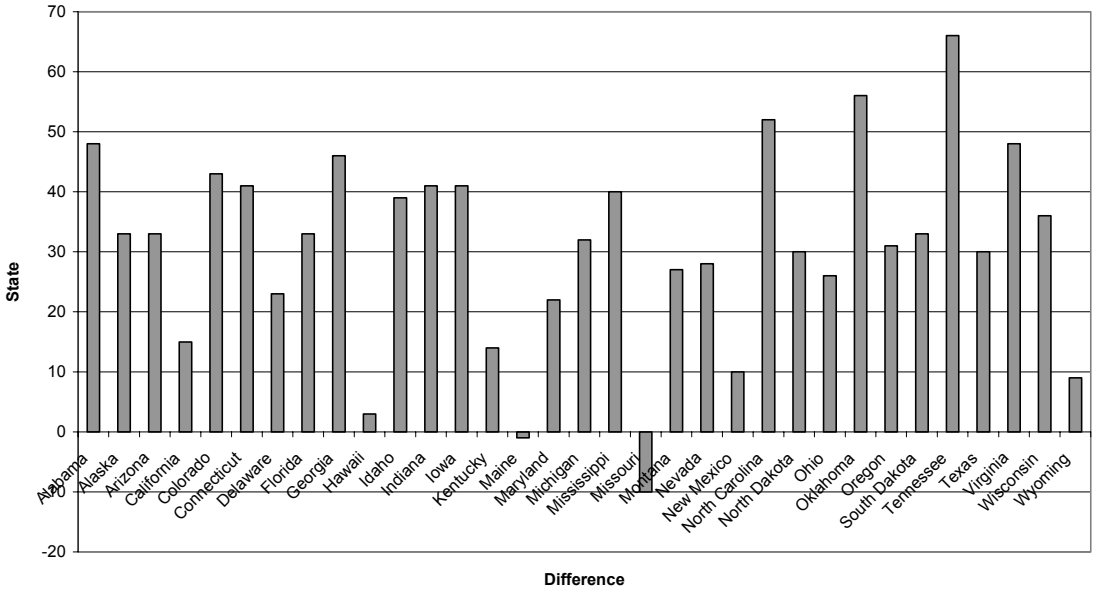

Figure 1. Difference between percentage of students proficient or above on 2005 state Grade 8 mathematics assessments and Grade 8 state NAEP mathematics assessments (33 states; source: Olsen, 2005).

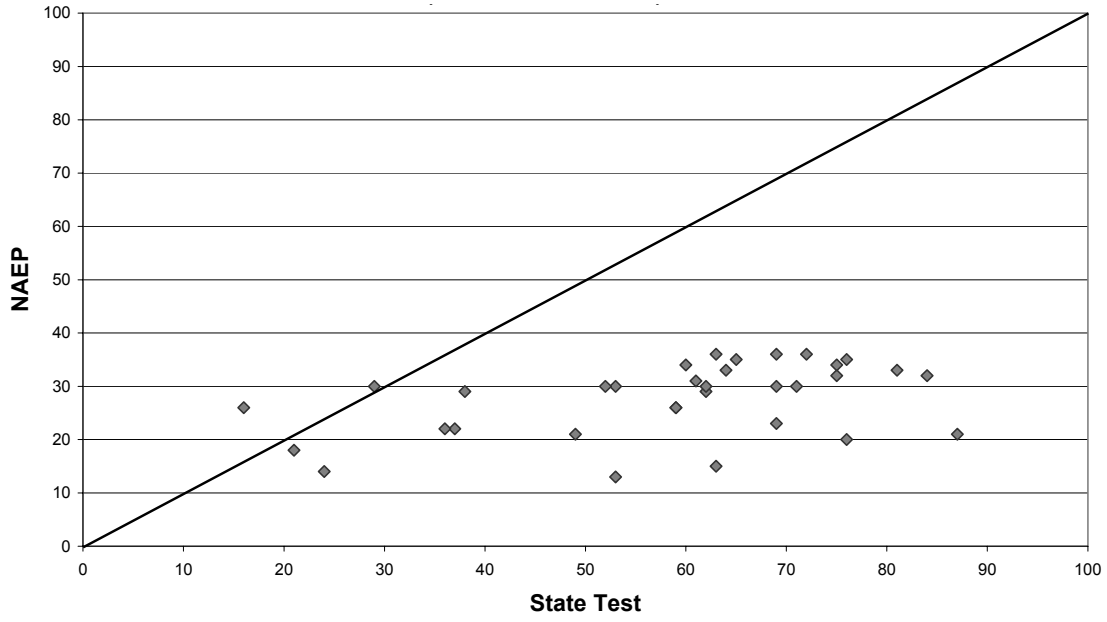

Figure 2. Scatterplot of percent proficient or above on Grade 8 state mathematics assessments and Grade 8 NAEP in 2005 for 33 states (r = .34; source: Olsen, 2005).

Alignment of Assessments with Content Standards

Academic content standards are supposed to specify what teachers should teach and students should learn. To accomplish these ends, the "Standards must be specific enough to enable everyone (students, parents, educators, policymakers, the public) to understand what students need to learn" (Linn & Herman, 1997, p. 1). Specificity is also required if the academic content standards are to serve as the basis for developing the specifications for assessments. Alignment of assessments with the content standards is critically important. Although the content standards are supposed to provide the targets for instruction, in practice, it is often the assessments that drive instruction. Thus, at the very least, assessments should reinforce the content standards rather than divert attention away from them and that requires a high degree of alignment.

Alignment has received considerable attention in the last few years, in part, because alignment of assessments and content standards is considered critical to the development of an effective standards-based accountability system and, in part, because NCLB requires states to provide evidence that their assessments are aligned with their academic content standards. Several different methods have been developed for evaluating alignment. Bhola, Impara, and Buckendahl (2003) have reviewed several of the more widely used methods. Although the methods vary in complexity and a number of specific details the more widely used methods (e.g., Porter; 2002; Webb, 1999) evaluate the match of the test's content to the standard, the match of the test to the range of knowledge and skills specified in the standards, whether there is a match in the balance of coverage, the match of the test to the level of cognitive demand called for in the standards, and whether the test includes material not found in the standards.

States must provide evidence regarding their standards and assessments for consideration as part of the NCLB peer review process. Part of this evidence must be addressed to the alignment of the state's assessments with their academic content and academic achievement standards. According the NCLB peer review guidance (U.S. Department of Education, 2004), evidence regarding alignment needs to be presented that the assessments:

> "Cover the full range of content specified in the State's academic content standards, meaning that all the standards are represented legitimately in the assessments; *and*

> Measure both the content (what students know) and the process (what students can do) aspects of the academic content standards; *and*
>
> Reflect the same degree and pattern of emphasis apparent in the academic content standards (e.g., if academic content standards place a lot of emphasis on operations then so should the assessments); *and*
>
> Reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the State's academic content standards; *and*
>
> Yield results that represent all achievement levels specified in the State's academic achievement standards" (U.S. Department of Education, 2004, p. 41, emphasis in original).

The peer review guidance clearly sets the bar at a high level with regard to the evidence that states need to provide to support the claim that their assessments are aligned with their academic content and achievement standards.

Test-Based Accountability Systems

Standards and tests may be the core of an accountability system, but accountability systems also require rules for the use of test results and usually specify a set of conditions for sanctions and, in some cases, rewards. There are two major approaches to the use of test results in accountability systems. The first is to compare test results at a given point in time to performance targets. Schools that meet or exceed the targets are considered to be successful, while schools where the test results fall short of the target are considered unsuccessful and may be subject to sanctions. This approach is sometimes referred to as a current status model. It has also has been called a school-mean performance approach (Raudenbush (2004a). The NCLB accountability system uses this current status approach. Schools meet adequate yearly progress (AYP) requirements if the percentage of students for the school as a whole and for each of several subgroups meet or exceed the annual performance targets in both reading/English language arts and mathematics.[1]

[1] The one exception to the comparison of current status to the annual targets is the "safe harbor" provision. The safe harbor provision allows a school that would not otherwise make AYP to do so if the percentage of students in the school as a whole and for each relevant subgroup scoring below proficient has decreased by at least 10% from the previous year in both reading/English language arts and mathematics and there is improvement on another performance indicator.

The second general approach to using test results in accountability systems is to evaluate growth in achievement. In the simplest form, the performance of successive cohorts of students in a school (e.g., fifth grade students in the current year vs. fifth graders the previous year) is compared. Kentucky's accountability system and California's Academic Performance Index (API) are examples of the successive cohorts approach to measuring improvement in student achievement. A more sophisticated approach uses growth in achievement of individual students from one year to the next. Growth models based on tracking score changes of individual students are much more demanding than the successive-cohort approach to measuring improvement, since they require the longitudinal tracking of individual students from year to year. North Carolina and Tennessee provide examples of states that use this longitudinal-tracking approach.

The matched longitudinal analysis of growth has substantial conceptual and practical advantages over the estimation of gains by the comparison of the performance of successive cohorts of students. There is also a great deal of uncertainty associated with the observed differences in scores of successive cohorts of students due to sampling error (Hill & DePascale, 2003; Linn & Haug, 2002; Linn, Baker & Betebenner, 2002; Zvock & Stevens, 2006). Interpretation of results from the successive cohorts approach also requires an assumption that the students attending the school are comparable from year to year in terms of background characteristics and prior learning. Changes in the composition of the student body due to student mobility call into question the basic assumption of comparability of the successive cohorts of students (Rouse, 2005). A growth model approach based on longitudinal tracking of individual students does not require the assumption that cohorts are comparable, but mobility can still cause problems if it leads to students being lost from the database (Rouse, 2005).

At a rhetorical level, the current-status approach has appeal because it sets the same performance expectations for all students and schools regardless of where students start. Thus, it avoids the pitfall of having higher standards for some students (e.g., affluent white students) than other students (e.g., poor minority students). On the other hand, current-status approaches, such as that used to determine AYP, "pose the greatest challenges to high-poverty schools, which enroll a large percentage of students who have traditionally scored poorly on standardized achievement tests (Kim & Sunderman, 2005, p. 4). As Kim and Sunderman go on to note, high poverty schools also enroll a disproportionate number of African

American or Latino students. Not surprisingly, they found that in each of the six states they studied, schools that according to NCLB were in need of improvement on average had larger percentages of African American or Latino students and smaller percentages of Asian American or white students than schools that made AYP (Kim & Sunderman, 2005, Table 1). Such differential impact on schools with larger numbers of African American and Latino students is consistent with the policy goal of closing the achievement gap, but still may be considered unfair to those schools where gains in achievement are being made, but just not fast enough to surpass the fixed performance targets for a given year.

A system based on growth in performance has the appeal that it attends to student learning and thereby may be seen as fairer to students and schools because it takes into account previous performance. But some advocates of poor and minority children worry that a growth model will result in lower expectations for "the nation's most disadvantaged young people" (Olson & Hoff, 2005, p. 16) by taking into account poor past achievement and thereby setting lower current achievement targets for those students. Thus, it is clear that there are advocates with opposing points of view for both a current-status and a growth-model approach to accountability.

Both current-status and growth-model approaches clearly are thought to have strengths and weaknesses. Thus, it is not surprising that several state accountability systems (e.g., Florida, Kentucky, Massachusetts) rely on a combination of status and growth to distinguish schools that are identified as successful from ones that are considered less successful.

The current-status and growth approaches, whether using successive cohorts of students or longitudinal tracking of students to compute gains yield different, often conflicting, results (Linn, 2005a; Raudenbush, 2005; Zvoch & Stevens, 2006). A school that meets the targets of a current-status accountability system may not show the gains required to be considered successful under a growth-model accountability system and vise versa. Hence, schools in states with either a successive cohorts improvement or a longitudinal growth model approach to school accountability may receive mixed messages from the state accountability and NCLB (Linn, 2005a). That is, a school may be considered successful or effective according to one accountability system but unsuccessful or in need of improvement by another system. Mixed messages of this kind are confusing to educators, parents, and the general public.

In response to interest expressed by several states in being able to use an approach that focuses on growth for purposes of NCLB, the U.S. Department of Education convened groups of experts in the summer of 2005 to consider ways in which growth models might be used for purposes of calculating AYP. On November 21, 2005, Secretary Spellings announced a pilot program that would allow states to submit proposals to use a growth model to make AYP determinations. In her letter to Chief State School Officers, Secretary Spellings listed 7 "core principles" that a growth model would have to meet to be approved. The first core principle specifies that the growth model "must ensure that all students are proficient by 2013-14 and set annual goals to ensure that the achievement gap is closing for all groups of students" (Spellings, 2005). Thus, the fixed achievement target of 100% proficient or above in 2013-2014 is maintained. Other core principles include the requirement of setting "high expectations for low achieving students, while not setting [them on] student demographic or school characteristics" and the requirement to make separate accountability decisions for reading/language arts and mathematics (Spellings, 2005).

Proposals to participate in the growth model pilot program were submitted by eight states. On May 17, 2006 Secretary Spellings announced approval for implementation of growth model pilots in 2005-2006 for two of those states, North Carolina and Tennessee. The other six states (Alaska, Arkansas, Arizona, Delaware, Florida, and Oregon) can have early consideration of revised proposals for possible implementation in 2006-2007.

The pilot program option opens the door for states with longitudinal tracking systems to propose models that may make their state accountability systems more compatible with the NCLB system. On the other hand, the pilot program does not provide a similar option for states using a successive-cohorts approach to measuring improvement, though as will be discussed below, one of the enclosures to Secretary Spellings' letter does describe index systems currently approved for use in nine states for calculating AYP that it might be argued could be used to track improvement in achievement of successive cohorts of students.

Illustrative Features of Accountability Systems

Compensatory and Multiple-Hurdle Approaches

The distinctions between accountability systems that focus on current status, those that focus on growth, and those that use some combination of current status

and growth are central considerations in determining what is valued by different approaches. There are several other features, however, that are also important in distinguishing accountability systems. Some systems allow for high performance in one subject area to compensate for lower performance in another area, while others treat each subject as a separate hurdle. The Kentucky accountability system, which has been in use for over a decade, uses the compensatory approach. An academic index is calculated based on test results at selected grades in seven content areas (reading, writing, mathematics, science, social studies, arts and humanities, and practical living/vocational studies) (http//www.kde.state.ky.us/). High student achievement in mathematics, for example, can compensate for somewhat lower performance in reading, or vise versa in the school's overall academic index score. California's API is also uses a compensatory approach to combine student achievement on tests in different subjects.

NCLB, on the other hand, uses a multiple-hurdle approach. Students must meet targets in both mathematics and reading/English language arts in order for the school to meet its AYP requirements. The school must also have at least 95% of the eligible students included in each assessment and meet another academic indicator defined by the state. Thus, even if there are no subgroups of students in the school with a sufficient number of students for disaggregated reporting, the school must clear all five hurdles to make AYP.

Although, as will be discussed below, the U.S. Department of Education has shown a willingness to be flexible regarding some aspects of AYP calculations, the multiple-hurdle approach is not one of those areas of flexibility. Indeed, in an enclosure to Secretary Spellings' November 21, 2005 letter discussing the growth-model pilot program explicitly indicates that states using assessments in other content areas in their model "should demonstrate that achievement on those other assessments does not compensate for low achievement in reading/language arts and mathematics" (Spellings, 2005). Nor can high achievement in reading compensate for lower achievement in mathematics.

Disaggregation

Systems differ in requirements for disaggregated reporting of results for subgroups of students and what the requirements are for subgroup performance on tests. California's API is a weighted combination of performance on tests in different content areas that is scaled to have a range of scores from 200 to 1,000. Eight

hundred has been established as the statewide API target. California sets a schoolwide growth target for a school that is equal to 5% of the difference between the school's API in the base year and the statewide target of 800. For example, a school with a base year API of 700 would have a growth target of 5 points (5% of 800—700) while a school with a base year API of 600 would have a growth target of 10 points. In addition to calculations of growth for the school as a whole, a school must show "comparable improvement" for numerically significant ethnic and economically disadvantaged subgroups of students. "Comparable improvement" is defined as 80% of the schoolwide annual growth target (California Department of Education, 2005).

The NCLB requirements for disaggregated reporting of subgroup reporting involves more subgroups and each subgroup with a sufficient number of students is used to define additional hurdles for making AYP using the same rules as are used for all students in a school. Disaggregated reporting is required for economically disadvantaged, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency.

For large schools with diverse student bodies, the disaggregated reporting requirements can greatly increase the number of hurdles a school must clear to make AYP. A school with a large enough number in, say, three racial/ethnic groups, students with limited English proficiency, economically disadvantaged students, and students with disabilities, would have a total 29 hurdles to clear, 4 for each of the 6 subgroups plus the 5 that all schools have for the total student body.

Disaggregated reporting is essential for monitoring the degree to which achievement gaps are closing. It is clear, however, that when combined with NCLB's multiple hurdles approach, disaggregation rules make it considerably more difficult for large schools with diverse student bodies to meet AYP requirements than it is for schools with homogenous student bodies (Kim & Sunderman, 2005; Linn, 2005a). As was previously noted, however, the greater challenge for schools with diverse student bodies is consistent with the policy goal of NCLB to close the gaps in achievement among racial/ethnic groups and between economically disadvantaged students and students coming from more affluent backgrounds.

The Department of Education has allowed states to introduce a number of refinements in their AYP determinations that mitigate the difficulties caused for schools with multiple subgroups of students to some extent. A report released on

November 16, 2005 by the Center on Education Policy (CEP) provides an update on the changes that states have made with the approval of the Department of Education that make it easier for schools meet AYP requirements. Several of the allowed changes help schools that have multiple subgroups. For example, more states are now using confidence intervals not only for comparing results to annual targets, but in safe harbor determinations. Confidence intervals lower the effective percent proficient target and the stringency of the safe harbor requirement, particularly for smaller subgroups, thereby making it easier for a school to make AYP (CEP, 2005; Porter, Linn, & Trimble, 2005).

Some states have been allowed to use performance indexes (see below) that give partial credit to students scoring below the proficient level. Retesting students with an alternate version of the test allows use of a student's best scores is also allowed in some instances. Finally, the minimum subgroup sizes have been increased which means that fewer subgroups have to meet AYP targets (CEP, 2005; Porter, Linn, &Trimble, 2005). Each of these changes makes it somewhat easier to meet AYP requirements.

Performance Indexes

Focusing only on the percentage of students who score at the proficient level or above has several potential shortcomings. It does not give credit for moving students from the lowest performance levels to higher levels of achievement that fall short of the minimum score required to be categorized as proficient. It has sometimes been suggested that this encourages teachers to focus their attention on "bubble students," that is, students who are performing near the proficient level, at the expense of more needy students who are performing so far below the proficient cut that there is little hope of raising their achievement enough to surpass the proficient cut score and at the expense of high achieving students for whom there is little doubt that they will score at the proficient level or above. By giving partial credit to students who are performing below the proficient cut, index scores avoid the disadvantages of only giving credit for students who are proficient or above.

Nine states (Massachusetts, Minnesota, Mississippi, New Mexico, New York, Oklahoma, Pennsylvania, South Carolina, and Vermont) have been approved to use index score in their AYP determinations (Spellings, 2005). For example, Massachusetts defined six levels of performance called advanced, proficient, needs improvement—high, needs improvement—low, warning/failing—high, and

warning/failing—low. Index scores are computed by Massachusetts by awarding 100 points for each student scoring at the proficient or advanced levels, 75 points for each student in the needs improvement—high category, 50 points for each student in the needs improvement—low category, 25 points for each student in the warning/failing—high category, and 0 points for each student in the warning/failing—low category (http://www.doe.mass.edu/sda/ayp/about.html?section=3).

In the baseline year, generally 2002, there was no advantage of using index scores rather than percentage of students at the proficient level or above, because the latter is used to define the starting level and the 2013-2014 target of 100% proficiency, which are then translated into the index score scale. In subsequent years, however, the index scores may show increases where the percent proficient or above is unchanged. This difference is illustrated for a hypothetical school in Table 1 using the Massachusetts index score system. As can be seen in Table 1, the number, and therefore the percentage because the total number of students is unchanged, of students who are at the proficient level or above is the same in 2006 and 2007. Hence, without an index score the school's rating in terms of percent proficient or above would the same in 2007 as it was in 2006. Because more students score in the higher achievement levels below the proficient cut in 2007 than in 2006, however, the index score shows a substantial improvement in 2007. This school clearly would benefit from the gains made in the percentage of students in the three higher categories below the proficient cut and would not be penalized for failing to also increase the number of students in the proficient or advanced categories of performance.

Table 1

Illustration of MA Index Scores for a Hypothetical School in 2006 and 2007

Performance Level	Points	N—2006	N - 2007	2006 Index Points	2006 Index Points
Prof. or adv.	100	50	50	5,000	5,000
NI—high	75	75	100	5,625	7,500
NI—low	50	100	125	5,000	6,250
W/F—high	25	100	125	2,500	3,125
W/F—low	0	75	50	0	0
Total		400	400	18,125	21,875

Prof. = proficient, Adv. = Advanced, NI = needs improvement, W/F = warning/failing, N = number of students.

2006 Index Score = 18,125/400 = 45.31

2007 Index Score = 21,875/400 = 54.69

Validity of Inferences from Accountability Systems

The simple claim that school A has met its AYP target requires only a modest inference that can be fairly easily validated. If the proportion of students in school A who perform at the proficient level or above exceeds the performance targets in both reading/English language arts and mathematics for the school as a whole and for each relevant subgroup of students, and the school has met other requirements such assessing at least 95% of the eligible students in each subgroup, then the validity of the AYP claim is buttressed. To complete the validity argument (Kane, in press), one would need to provide evidence that the tests are measuring what they are supposed to measure, are adequately aligned with the content standards, for example. In addition, assurance would be needed that the tests were appropriately administered and that scores were not inflated due to cheating. Overall, however, the validation of the claim is rather straight forward.

Causal Interpretations

Now consider the claim that school A is successful or that it is more effective than school B, where the proportion of students scoring at the proficient level or above was less than the AYP target. These claims are much more difficult to justify from the information provided by the NCLB accountability system. Indeed, Raudenbush (2004a) has made a convincing case that such claims are "scientifically indefensible" (p. 35).

As Raudenbush has argued, such a claim that school A is more effective than school B requires an inference that the schools and their instructional programs that have caused the better achievement in school A than in school B. Such a causal inference, however, requires that many other competing hypotheses be eliminated as alternative explanations of the differences in achievement of students in the two schools. The alternate explanation that students in school A were better readers and knew more mathematics than students in school B before the start of the school year, for example, provides an alternative explanation that needs to be eliminated before concluding that school A is more effective than school B.

Many other alternate explanations, such as differences in educational support from home for students in schools A and B or differences in composition effects created by the peer groups in the two schools, would also have to be ruled out. School characteristics are confounded with many factors (e.g., socio-economic status and prior achievement of students). According to Myers (2000), for example, current status school accountability measures 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" (p. 2). Consequently, differences in achievement at a fixed point in time simply do not provide a defensible justification for the causal inference that school A is more effective than school B.

Accountability systems that focus on year-to-year change can rule out, or at least make less plausible, some of the alternative explanations that make causal inferences impossible to defend in a current status approach to accountability. The longitudinal tracking of individual students, for example, makes it possible to eliminate the explanation that greater gains in school A than in school B, are due to differences in the achievement of students in two schools at the beginning of the year. This is not so straight forward for successive-cohorts approach to measuring gains because, as was noted above, the students in a given grade the previous year may not have been comparable to their counterparts in the current year.

Average gains in achievement based on longitudinal tracking of individual students provide a stronger basis for eliminating competing explanations for differences in school performance. Sophisticated value-added statistical models (e.g., Sanders & Horn, 1998; see also, Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004) help in ruling out the possibility that

initial differences in student achievement rather than differences in school instructional programs explain the better gains in one school than another. The "value-added" terminology implies a causal interpretation. When teacher or school value-added results are reported it is assumed that it is the teacher or the school that is having an effect rather than some other factor such as students' families, student background, or student peers in the school.

Most applications of value-added models estimate gains based on a span of grades, the earliest of which is likely to be grade 2 or 3. In this way the value-added models control for differences in student achievement at the time of the earliest grade included in the analysis, but they do not rule out the possibility that achievement differences in kindergarten and grade 1 confound the value-added estimates. As Raudenbush (2004a) has noted, "measured cognitive status prior to school entry is the most important confounder in studying school effects" (p.13).

Value-added models that use prior student achievement but do not include other student background are also subject to the criticism that the excluded variables might bias the estimates. In this regard, Raudenbush (2004a) has argued that,

> "… the estimation of gains does not necessarily eliminate all confounding. A critic might argue that unmeasured student characteristics predict gains students can expect and the schools they attend. This criticism is impossible to refute, though Ballou, Sanders, and Wright (2004) provide evidence that use of longitudinal data in multiple subject areas virtually eliminates the need to control for the usual confounders (ethnicity, gender, and poverty)" (p. 13).

The composition of the student body may influence both status and growth. Hanushek, Kain, Markman, & Rivkin (2003), for example, found that peer average achievement had a significant effect on the gains in achievement throughout the achievement distribution. The peer group is a characteristic of a school, but as Rouse (2005) has noted, it is not a characteristic that is under the control of public schools. Peer group effects, like school effects are difficult to isolate, but the peer group composition provides an alternative explanation that is not easily dismissed in attempts to attribute absolute performance or gains in achievement to school quality. Thus, while value-added or other types of growth models provide substantial improvements over the current-status and improvement for successive-cohorts approaches to accountability, they still fall short of providing definitive evidence that school differences in student gains in achievement are attributable solely to differences in school quality.

Causal claims are best supported when random assignment can be used because random assignment makes it possible to eliminate many alternative explanations of results. Since neither school practices nor students are assigned to schools at random, it is much harder to support causal claims based on test results obtained as part of accountability systems. Recognizing the difficulties in supporting causal claims, Rubin, Stuart, and Zanutto (2004) have argued that value-added analyses "should not be seen as estimating causal effects of teachers or schools, but rather as descriptive measures" (p. 113).

Similarly, Raudenbush (2004b) has concluded that "they should not be taken as direct evidence of the effects of instructional practice" (p. 128). In his rejoinder Ballou (2004) stressed the superiority of value-added models to the status accountability models such as that used for NCLB where schools are held accountable by comparisons of student performance "against the same absolute yardstick" (p. 133). As was already indicated, the value-added approach has substantial advantages over a current-status approach, but it is still found wanting when it comes to making casual claims that are required to reach conclusions about school quality—which is not only the goal, but a logical requirement for accountability systems that provide sanctions to "failing" schools or rewards to successful or exemplary schools if the systems are to be considered fair and valid.

Score Inflation

After a test-based accountability system has been implemented, whether the system emphasizes current status or growth, it is common to see annual increases in student achievement test results for the state as a whole. In Colorado, for example, the percentage of 4th grade students scoring at the proficient level or above on the reading assessments increased from 60% in 2000 to 64% in 2005 and the percentage for 8th grade students went from 35% in 2000 to 44% in 2005 on the mathematics assessments (Results obtained from http://www.cde.state.co.us/cdeassess/csap/as_latestCSAP.htm.) Similar increases were experienced with a number of other state testing programs. It is common and natural to infer that the increases in test scores reflect real improvement in student achievement, not just gains in test scores.

The validity of the inference that score gains reflect real improvement has been called into question, however, by a number of authors (e.g., Hamilton, 2003; Koretz, 2005). *Score inflation,* which Koretz (2005) defines as "as a gain in scores that

substantially overstates the improvement in learning it implies" (p. 99) is a major threat to valid interpretations gains in test scores. A number of studies have found that gains in scores on high-stakes tests often fail to generalize to other indicators of student achievement such as results on the National Assessment of Educational Progress (NAEP) or the ACT (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998).

Koretz, Linn, Dunbar and Shepard (1991) studied a district where sharp gains in test scores were reported for the first several years that a test was in place, followed by a large drop in scores when a new test was introduced after which steady gains were once again obtained. This saw tooth pattern had been frequently observed in similar circumstances (Linn, Graue, & Sanders, 1990), and was not unique to the district in the Koretz, Linn, Dunbar, and Shepard (1991) study. When Koretz, et al. (1991) re-administered the original test to students in the district, they found that the students scored at about the same level as they had the first year the test was used—well below the level that was achieved the last year the original test was administered and also well below the level achieved on the district's new test in the year when the original test was re-administered as part of the study (see Koretz, 2005, for a more complete description of the study and a graphical presentation of the results).

School Characteristics and Instructional Practices

A major shortcoming of current accountability systems for purposes of making valid inferences about school quality is due to severe limitations on the data that are generally included in the systems. Accountability systems now being used to meet state and federal requirements throughout the country usually lack information about instructional practices, teacher characteristics, and student characteristics other than student test scores and some student demographic data such as gender, race/ethnicity, indicators of economic disadvantage, disability status, and English language proficiency.

Student outcomes, even when prior achievement and student demographic information is included in the accountability analyses do not provide a sufficient basis for making the type of causal inferences that are implied when accountability system results are used to identify successful and unsuccessful schools and to impose sanctions on the latter schools. As Raudenbush (2004a) has argued, "to be successful, accountability must be informed by other sources of information, and, in

particular, information on organizational and instructional practice" (p. 37). Even with such additional information, causal interpretations will be difficult to justify and subject to challenge, but they would be on much firmer ground than is possible without the additional information about school organization and instructional practice in the schools being held accountable.

In an earlier era, resources and process information were used to draw conclusions about school quality. The current emphasis on outcomes is clearly preferable to a system that uses resources as the primary basis of judging school quality. It would be better still to base judgments about school quality on a combination of information about student outcomes, prior achievement and backgrounds of students, and school organizational and instructional processes.

Descriptive Uses of Accountability System Results

Accountability system results can have value without making causal inferences about school quality solely from the results student achievement measures and demographic characteristics. Treating the results as descriptive information and for identification of schools that require more intensive investigation of organizational and instructional process characteristics could be of considerable value. Rather than using the results of the accountability system as the sole determiner of sanctions for schools, they would be used to flag schools that needed more intensive investigation to reach sound conclusions about needed improvements or judgments about quality.

Such a use of accountability system results would represent a profound change and would require revisions of state and federal laws. It is unlikely that such a change in perspective would be politically acceptable at the present time. The change, however, would make the use of accountability results more consistent with the tenets of scientific reasoning and research. In this sense, it would make accountability requirements more consistent with the major emphasis on the importance of using "scientifically based research" that is found throughout the NCLB Act (Feuer, Towne, & Shavelson, 2002).

Conclusion

There are many reasons for the long-standing emphasis on test-based accountability in the United States. Accountability is a politically attractive means of trying to reform education. Compared to other reforms, it is relatively inexpensive and represents one of the few ways in which politicians and policymakers can have

an effect on classroom practice (Linn, 2000). Test results are used both to monitor changes in student achievement and gaps in the achievement different subgroups. Test-based accountability is also seen as an engine of educational reform that can both lead to improvements in student achievement and the closing of gaps in achievement among racial/ethnic groups and between economically disadvantaged students and their more affluent counterparts.

For at least the last decade, test-based accountability systems have been based on state-adopted content standards and results have generally been reported in terms of a small number of performance standards. The content standards specify what students are supposed to learn and they also provide the basis for developing tests that are consistent with and reinforce those standards. Performance standards specify the levels of achievement that are expected to be considered proficient or advanced. The passage of NCLB has given increased importance to a single performance standard with the great emphasis on the proficient level that all students are expected to achieve by 2013-2014.

There are a variety of approaches to test-based accountability. Systems can be distinguished in terms of their reliance on measures of current status in relationship to fixed targets or growth in achievement. NCLB uses a current-status approach while a number of state systems give greater priority to growth or use a combination of current-status and growth targets. Although accountability systems are used to make causal interpretations about school quality, such interpretations cannot be defended on scientific grounds. There are too many plausible alternate explanations of results to conclude that observed differences in the achievement of students in different schools are solely attributable to variations in school quality. Growth-model approaches are helpful in eliminating some, but not all plausible alternative explanations.

Accountability results are best viewed as a source of descriptive information about schools. They can be a source of hypotheses that need to be checked by gathering additional information about school organizational characteristics, teacher characteristics and instructional practice.

References

Ballou, 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*(3), 21-29.

California Department of Education. (2005, October). *2004-05 Academic performance index growth report: Information guide.* Sacramento, CA: Author. Available at http://www.cde.ca.gov/ta/ac/ap/documents/infoguide05g.pdf

Center on Education Policy. (2005). *States test limits of federal AYP flexibility*. Washington, DC: Author. Available at http://www.ctredpol.org/

Education Week. (2005, January 5). Quality Counts 2005: No small change: Targeting money toward student performance. *Education Week, 24*(17).

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4-14.

Goals 2000: Educate America Act of 1994, Pub. L. No. 103-227 (1994).

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & 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 I, pp. 1-34). Boston, MA: Blackwell Publishing.

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

Hanushek, E. A., Kain, J. F., Harkman, J. M., & Rivkin, S. G. (2005). Does peer ability affect student achievement? *Journal of Applied Economics, 18*(5), 527-544.

Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice, 22*(3), 12-20.

Improving America's Schools Act of 1994, Pub. L. No. 103-382 (1994).

Kane, M. (In press). *Validation*. In R. L. Brennan (Ed.), Educational measurement (4th ed.). Westport, CT: American Council on Education/Praeger.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3-13.

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. (2005). *Alignment, high stakes, and the inflation of test scores*. In J. L. Herman & 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 I, pp. 99-118). Boston, MA: Blackwell Publishing.

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

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

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

Linn, R. L. (2003, September 1). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved September 1, 2003 from http://epaa.asu.edu/epaa/v11n31/

Linn, R. L. (2005a, June 28). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Educational Policy Analysis Archives, 13*(33). Retrieved June 30, 2005 from http://epaa.asu.edu/epaa/v13n33/.

Linn, R. L. (2005b). Issues in the design of accountability systems. In J. L. Herman & 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 I, pp. 78-98). Boston, MA: Blackwell Publishing.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

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

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

Linn, R. L., & Herman, J. L. (1997). *A policymaker's guide to standards and assessment.* Denver, CO: Education Commission of the States. Available at http://www.ecs.org/ecsmain.asp?page=/html/publications/home_publications.asp.

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

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

McDonnell, L. M. (2005). Assessment and accountability from the policymakers' perspective. In J. L. Herman & 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 I, pp. 35-54). Boston, MA: Blackwell Publishing.

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.

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

No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2002).

Olson, L. (2002, January 9). Testing systems in most states not ESEA-ready. *Education Week, 31*(16), 1, 26-27.

Olsen, L. (2005, September 2). Defying predictions, state trends prove mixed on schools making NCLB targets. *Education Week, 25*(2), 1, 26-27.

Olsen, L. & Hoff, D. J. (2005). U.S. to pilot new gauge of growth: Ed. Dept. to permit shifts in how states track gains. *Education Week, 25*(13), 1, 16.

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

Porter, A. C., Linn, R. L., & Trimble, S. (2005). The effects of state decisions about NCLB Adequate Yearly Progress targets. *Educational Measurement: Issues and Practice, 24*(4), 32-39.

Raudenbush, S. W. (2004a). *Schooling, statistics, and poverty: Can we measure school improvement? The ninth annual William H Angoff Memorial Lecture*. Princeton, NJ: Educational Testing Service.

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

Rouse, C. E. (2005). Accounting for schools: Economic issues in measuring school quality. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 275-298). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Spellings, M. (November 21, 2005). Letter to Chief State School Officers, announcing growth model pilot program, with enclosures. Available at: http://www.ed.gov/policy/elsec/guid/secletter/051121.html

U.S. Department of Education. (2004). Standards and assessments peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001, April 28. Washington, DC: Author.

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

The White House. (2001). Press conference with President George W. Bush and Education Secretary Rod Paige to Introduce the President's Education Program, Office of the Press Secretary, January 23. http://www.whitehouse.gov/news/releases/2001/01/20010123-2.html

Zvoch, K. & Stevens, J. J. (2006, January 20). Successive student cohorts and longitudinal growth models: An investigation of elementary school mathematics performance. *Education Policy Analysis Archives, 14*(2). Retrieved January 20, 2006 from http://epaa.asu.edu/epaa/v14n2/