Test-based Educational Accountability in the Era of No Child Left Behind

CSE Report 651

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.

TEST-BASED EDUCATIONAL ACCOUNTABILITY IN THE ERA OF NO CHILD LEFT BEHIND

Robert L. Linn
CRESST/University of Colorado at Boulder

Abstract

The ever-increasing reliance on student performance on tests as a way of holding schools and educators accountable is discussed. Comparisons are made between state accountability requirements and the accountability requirements of the No Child Left Behind (NCLB) Act of 2001. The resulting mixed messages being given by the two systems are discussed. Features of NCLB accountability and state accountability systems that contribute to the identification of a school as meeting goals according to NCLB but failing to do so according to the state accountability system, or vise versa, are discussed. These include the multiple hurdles of NCLB, the comparison of performance against a fixed target rather than changes in achievement, and the definition of performance goals. Some suggestions are provided for improving the NCLB accountability system.

The assessment of student achievement has long been an integral part of education. Test results for individual students have been used for myriad purposes, such as monitoring progress, assigning grades, placement, college admissions, and in grade-to-grade promotion, and high school graduation decisions. The use of student test results to judge programs and schools, with a few exceptions (see, for example, Resnick, 1982), has a shorter, but still substantial, history. Both states and the federal government have moved away from resource and process measures as a means of judging the quality of schools to an ever-increasing reliance on student test results to hold schools accountable. The characteristics of the school accountability systems evolved over the last 40 years and the systems vary a good deal from one state to another, as do the state and federal accountability systems.

State Testing Initiatives

In the 1960s, policymakers in several states tried to introduce statewide testing programs that would provide a means of monitoring education and evaluating the effectiveness of schools. The attempts to use tests for school accountability generally failed to be approved, however. Instead, the statewide testing programs that were put in place in the 1960s were usually intended to be used for student guidance and the identification of talent (Mazzeo, 2001). But the push for accountability was not to be denied for long. In the following two decades, states moved toward statewide testing programs that had higher stakes initially for students and later for schools and educators.

Two-thirds of the states introduced some form of minimum competency testing during the 1970s and early 1980s (Office of Technology Assessment, 1992). "Minimum competency tests, which coincided with the 'back to the basics' movement of the 1970s, typically tested students on basic literacy and numeracy skills" (Heubert & Hauser, 1999, p. 38). During this same period, many states also introduced some form of annual statewide testing program for monitoring educational progress. The tests were generally based on statewide samples involving students at two or three grade levels (Anderson, 1982; Pipho, 1978).

Early Federal Requirements

The federal involvement in test-based accountability began in a significant way with the enactment of the Elementary and Secondary Education Act (ESEA) of 1965. Title I of ESEA provided financial support for compensatory education to schools serving poor children. Testing requirements for Title I students were instituted as the result of congressional demands for evaluation and accountability. Initially, the test requirements allowed schools to administer standardized, norm-referenced achievement tests to Title I students and results were generally reported in terms of grade-equivalent scores. It soon became evident that grade-equivalent scales varied greatly from one test publisher to another and from one content area to another, making comparisons across school districts or across states impossible (for discussions of properties of grade-equivalent scores, see, for example, Angoff, 1971; Linn & Slinde, 1977; Petersen, Kolen, & Hoover, 1989).

In an effort to cope with these problems, the U.S. Department of Education introduced the Title I Evaluation and Reporting System (TIERS) (Tallmadge &

Wood, 1981). TIERS encouraged the administration of norm-referenced standardized tests to Title I students in both the fall and the spring. Programs were evaluated in terms of normal-curve equivalent scores, which are normalized conversions of the publisher's percentile ranks obtained from their norms. The best estimates from the era of the 1970s and early 1980s were that typical gains were on the order of magnitude of between a tenth and a twentieth of a standard deviation per year (Linn, Dunbar, Harnisch, & Hastings, 1982). Fall to spring gains were typically followed by some loss in scores over the summer months, however (Hemenway, Wang, Kenoyer, Hoepfner, Bear, & Smith, 1978; Linn, et al., 1982). Although little use was made of the aggregate test results, these TIERS requirements relieved the pressure from demands for accountability for Title I.

Evolution of State Testing Requirements

The federal requirements for testing of Title I students certainly had an impact on the testing industry, but because the requirements were limited to students participating in Title I programs and because so little use was made of the results, the relevance for states and districts was limited. Many school districts and quite a few states introduced their own testing requirements in the 1970s and 1980s. Minimum-competency tests were introduced as high school graduation requirements and norm-referenced standardized tests were administered at selected grades throughout the country. Not surprisingly, there was a rapid growth in the sales of both off-the-shelf and customized standardized tests. The expenditures on standardized tests in constant 1998 dollars more than tripled between 1977 and 1989 (Clarke, Madaus, Horn & Ramos, 2001).

Pipho (1984) reported that 40 of the 50 states had some form of competency testing. "Of the 40 states with some competency testing, statewide standards [were] imposed in 21 states, standards [were] determined by local school systems in 10 states, standards [were] determined by a combination in 7 states, and no uniform standards [were] imposed in 2 states" (Jaeger, 1989, p. 489).

The standards-based assessment movement swept the country during the 1990s. The standards-based reform movement was predicated on the idea, articulated by Smith and O'Day (1990), that student performance would improve if states adopted content standards and statewide assessments that were intended to measure those standards. By the time the No Child Left Behind (NCLB) Act of 2001 (Public Law 107-110) law was signed by President Bush in January 2002, every state

except Iowa had adopted content standards and was using assessments for some form of school accountability. The exception of Iowa is rather ironic since almost all schools in Iowa have administered the Iowa Test of Basic Skills or the Iowa Tests of Educational Development every fall for decades.

Mixed Messages

States use different assessments, have adopted their own student performance stands, and have developed different accountability systems. As a consequence of those differences, reports of student achievement and progress are not comparable from one state to another. Although the requirements of NCLB have led to greater commonality among states in some respects, a great deal of between-state variability remains in many important details in implementing assessments and accountability provisions.

NCLB requires states to test students in Grades 3 through 8 in mathematics and English/language arts starting no later than the 2005-'06 school year. NCLB requires each state to have adopted "challenging academic content standards and challenging student academic achievement standards" (P. L. 107-110, Section 1111(b)(1)(A)). States must also establish adequate yearly progress (AYP) goals for each year from 2002 to 2014 that culminate in the 2014 goal where all students are at or above the proficient student academic achievement standard. As discussed below, however, states still control many aspects in complying with NCLB, such as the specification of content standards, the choice of assessments, and the setting of academic achievement standards.

For states with functioning assessment and accountability systems of their own, NCLB accountability has frequently been layered on as a separate system. Kentucky, for example, has had an accountability system in place for several years that uses tests in seven content areas (reading, writing, mathematics, science, social studies, arts and humanities, and practical living/vocational studies). The tests are administered at selected grades so that the overall testing burden is distributed across grade levels. Composite index scores that are derived across content areas and which include some non-test measures (e.g., attendance or graduation rate) are used for school accountability. Biennial accountability targets for the composite index scores are set for schools relative to the schools' starting position defined by the school's accountability index score in the 1999-2000 biennium. Schools that started low have to gain more in their index score than schools that started out

relatively high, but all are supposed to reach an index value of 100 by the 2013-2014 biennium (Kentucky Department of Education, 2004). In the computation of the index value, students who score at the highest level (called distinguished) on a test contribute 140 points; students at the proficient level contribute 100 points; and students in various categories below proficient contribute an amount less than 100—how much less depends on how far below the proficient level the student's score is.

NCLB imposes a quite different set of accountability requirements for Kentucky schools. Mathematics and reading must be reported separately and schools must make annual, rather than biennial, measurable objectives in each subject (not just on a composite score). No extra credit is allowed for students scoring at the distinguished level. They are simply lumped with proficient students in the proficient or above category. School performance is compared to an absolute target, which is the same regardless of where the school started. Schools must meet AYP requirements in both reading and mathematics, not only for the student body as a whole, but for each of several subgroups (assuming there are enough students in a subgroup to be counted for NCLB accountability purposes): major racial and ethnic groups, English proficiency status, migrant status, student disability status, and economic status.

Given the differences between Kentucky's own school accountability system and the NCLB system, it is hardly surprising that the two systems are giving mixed messages. Of Kentucky schools, 730 of 986 (74.0%) made AYP in 2004. According to Kentucky's own accountability, however, 943 of the 986 (95.6%) schools met their Kentucky biennium goals in 2003-2004. [1] Thus, the best possible agreement between the two systems would be if all of the 730 schools that made AYP also met the Kentucky biennium goals also made AYP and all of the 43 schools that fell short of the Kentucky biennium goals also failed to make AYP for a combined total agreement of the two systems for 773 of the 986 (78.3%) schools. Even in this best-case scenario, just over 20% of the schools would receive mixed messages by meeting the goal according to one accountability system, but failing to do so according to the other system.

Table 1 displays the cross-classification of the 986 schools in terms of meeting or not meeting AYP in 2004 and meeting or not meeting Kentucky's accountability goals. As can be seen, not all the schools that made AYP also met the Kentucky

[1]Excludes schools with special designations. Based on Ford and Thacker (2005).

Table 1

Cross-tabulation of the state and NCLB classifications of Kentucky schools in 2004 number of schools (percentages of all schools in parentheses)*

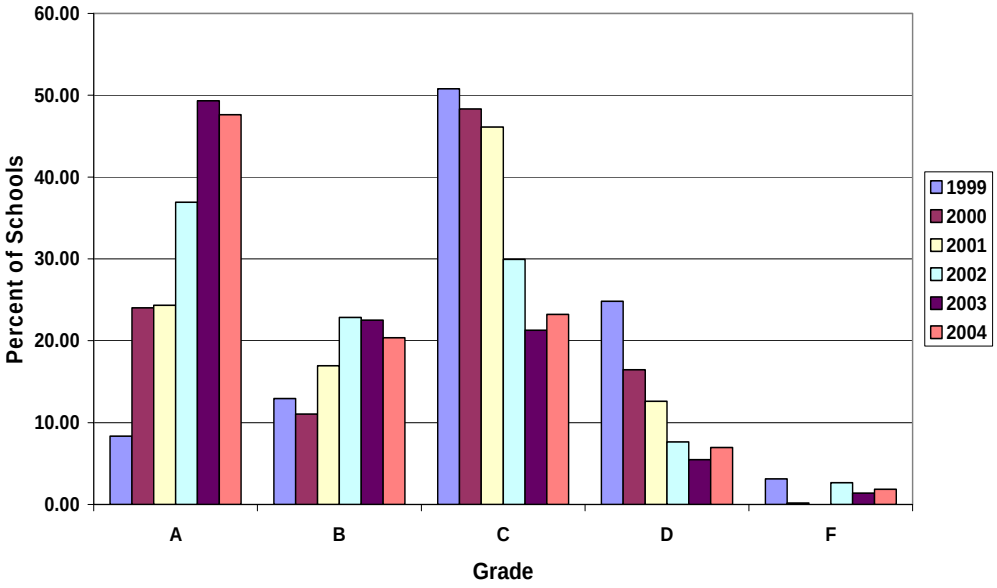
Met the state goal	Met AYP targets		
	No	Yes	Total
Yes	230 (23.3)	713 (72.3)	43 (4.4)
No	26 (2.6)	17 (1.7)	943 (95.6)
Total	256 (26.0)	730 (74.0)	986 (100)

* Based on Ford and Thacker (2005).

goals. Hence, a quarter of the schools (247 of 986) received mixed messages that they met expectations according to one accountability system but failed to meet them according to the other system.

The mixed messages of the NCLB and individual state's own accountability systems are not unique to Kentucky. Florida's state accountability system has assigned letter grades of A, B, C, D, or F to schools based on the performance of students on their state assessment. The distributions of Florida school grades over the past 6 years are displayed in Figure 1. These distributions have painted quite a favorable picture of school performance. The percentage of schools receiving grades of A has increased from 8.3% in 1999 to 47.6% in 2004. The percentage of schools receiving either an A or a B has also increased sharply (from 21.3% in 1999 to 68.0% in 2004, while the percentage of schools receiving Ds or Fs has declined from over a quarter of the schools in 1999 (27.9%) to less than a tenth of the schools in 2004 (8.8%).

Figure 1
Distribution of Florida School Grades by Year
(Source: http://schoolgrades.fldoe.org/o304sg-page01.pdf)



The NCLB accountability results in Florida in the last 2 years have provided a sharp contrast to the positive results from Florida's own accountability system. In 2003, Florida had the dubious distinction of leading the nation as the state with the largest percentage of schools (82%) that failed to make AYP. Although there was some decline in the percentage of Florida schools that did not make AYP in 2004 (from 82 to 77%), only Alabama had an equally high percentage of schools failing to make the AYP target in 2004 (Olson, 2004, p. S6). Other southern states had more modest percentages of schools that failed to make AYP in 2004: Georgia, 20%; Louisiana, 8%; Mississippi, 24%; North Carolina, 29%; South Carolina, 44%, Tennessee, 14%; and Virginia, 25% (Olson, 2004, p. S6). As will be discussed in greater detail below, the variation by state makes little sense in comparison to other information about student performance by state, such as that provided by the National Assessment of Educational Progress (NAEP), but it is nonetheless clear that the state's own accountability system and NCLB are giving quite a mixed picture in Florida. Fifty six percent of the 1,262 schools in Florida that received an A in 2004 failed to make AYP.

The mixed messages provided in Kentucky and Florida are repeated in varying degrees in many other states. Colorado, for example, has an academic performance

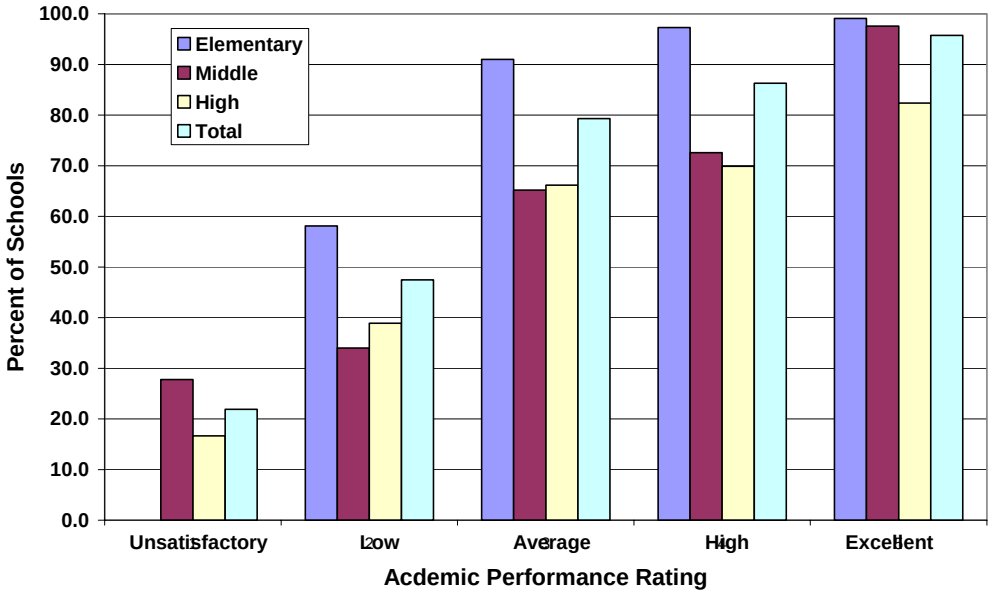
rating system that assigns schools to one of five graded performance categories called, Unsatisfactory, Low, Average, High, and Excellent. Figure 2 displays the percentage of schools that made AYP in 2003 by school type and academic performance rating. As can be seen in Figure 2, there is a clear relationship between the Colorado academic performance rating of a school and the likelihood that the school will meet AYP. The relationship is far enough from perfect, however, to result in mixed messages for a substantial number of schools. For example, 21.9% of the schools rated "Unsatisfactory" and 47.5% of the schools rated "Low" made AYP, while 13.7% of the schools rated "high" failed to make AYP.

Although the summary statistics provide clear evidence that mixed messages are being given by NCLB and state accountability systems, as was clearly illustrated by Dillon (2004), it is the failure of prestigious suburban high schools to meet AYP requirements that seems to have caused the most consternation. Dillon quotes people such as Representative Judy Baggert of Illinois, who "helped write the law" and former North Carolina Gov. James Hunt, who has praised the law, among others who were dismayed when they learned that particular high schools that they knew to be excellent were identified for failing to make AYP. Although as is explained below there are a variety of aspects of the NCLB identification system that makes it likely that excellent schools will be found wanting by failing to make AYP it is nonetheless confusing to parents and the general public. The confusion is summed up well in the title of Dillon's article "Good schools or bad? Conflicting ratings leave parents baffled" (2004).

Accountability a la NCLB

There are several features of the NCLB accountability requirements that make it likely that the results will conflict with the accountability of individual states. Some of these features also contribute to the wide state-to-state disparities in the proportion of schools that meet AYP. Three of these features, the use of absolute targets rather than improvement targets relative to a school's starting level, the need to meet targets in both reading and mathematics rather than a composite, and the requirement of meeting targets for separate subgroups within a school, were mentioned in passing in the discussion of the Kentucky results. These and other features are elaborated in this section.

Figure 2
Percentage of Colorado Schools Making AYP in 2003 by School Type and Academic Performance Rating
(Source: http://www.ced.state.co.us/)



Multiple Hurdles

Unlike many state accountability systems, NCLB requires schools to clear several hurdles. The most obvious of these is that student achievement must exceed the annual measurable objective (AMO) in both mathematics and reading/English language arts. Performance on, say, reading achievement that far exceeds the AMO cannot compensate for mathematics achievement that just misses the AMO. In addition to meeting the separate AMOs for mathematics and reading/language arts, schools must have at least 95% of their eligible students participate in the assessments in each subject. The school must also meet the goal established for the "other academic indicator," usually attendance rate for elementary and secondary schools and graduation rate for high schools, required by NCLB. Thus, there are a minimum of five hurdles for a school with a homogeneous student body and insufficient numbers in any subgroup to be held accountable for disaggregated results.

The number of hurdles for meeting AYP expands rapidly for large schools with diverse student bodies due to the disaggregation requirements of NCLB. A school

with more than the minimum number of students (designated by the state and approved by the U.S. Department of Education) for purposes of AYP in each of the following subgroups—4 racial ethnic groups, students with limited English proficiency, economically disadvantaged students, and students with disabilities—would have not 5, but 33, hurdles to clear (the 5 when all students in the school are considered as a whole, plus 16 for the 4 hurdles for each of 4 racial/ethnic groups, plus 4 for students with limited English proficiency, plus 4 for the economically disadvantaged students, plus 4 for the students with disabilities [see Table 2]).

Thus, schools can meet AYP requirements in only one way, by clearing multiple hurdles, but can fall short in many different ways. Given the larger number of hurdles to be cleared by more diverse schools, it is not surprising that Novak and Fuller (2003) found that schools serving more diverse student bodies were less likely to meet AYP requirements than schools serving less diverse student populations.

Table 2

Illustration of NCLB's multiple hurdles for a large school with a diverse student body*

	Reading/English language arts		Mathematics		
Group of students	Participation rate	Percent proficient or above	Participation rate	Percent proficient or above	Other academic indicators
All students	1	2	3	4	5
Racial ethnic group 1	6	7	8	9	
Racial ethnic group 2	10	11	12	13	
Racial ethnic group 3	14	15	16	17	
Racial ethnic group 4	18	19	20	21	
Economically disadvantaged	22	23	24	25	
Students with limited English proficiency	26	27	28	29	
Students with disabilities	30	31	32	33	

*Table modeled after Marion, White, Carlson, Erpenbach, Rabinowitz, and Sheinker (2002).

"Even when students display almost identical average test scores schools with more subgroups are more likely to miss their growth targets under federal rules set by the No Child Left Behind Act" (Novak & Fuller, 2003, p. 1). These results are to be expected because even a school with high average achievement may be tripped up on one of the multiple hurdles, such as missing the participation rate for a particular subgroup, or because students with disabilities perform below the proficient cutoff in either reading or mathematics. "In Westport, Conn., (for example) the Bedford Middle School, where test scores are often among Connecticut's highest, was called low performing because the school failed to meet the 95% standard for testing for the disabled by one student" (Dillon, 2004).

Status vs. Growth Targets

State accountability systems frequently establish performance targets based on growth, thereby taking into account previous performance as well as current status. NCLB requirements, on the other hand, with the exception of the safe harbor provision discussed below, focus only on current status in comparison to the performance target. California's accountability system, like the Kentucky system described previously, provides a good example of a carefully developed system that focuses on growth in achievement. California's system uses an academic performance index (API) that is a weighted combination of performance on tests of English language arts (including writing) and mathematics for Grades 2 through 8. For Grades 9 through 11, history-social science and science are included along with English language arts and mathematics for the weighted API. The API is scaled to have scores that range from 0 to 1,000.

An API score of 800 has been selected by the State Board of Education "as the target toward which all schools should aspire" (California Department of Education, 2004, p. 29). Schools are not sanctioned for falling short of the absolute target of 800, however. Instead a school is held accountable for meeting their annual API growth target, which "is defined as 5% of the distance from the school's API and the statewide performance target or a minimum of one point" (California Department of Education, p. 30). For example, a school with an API in 2003 of 700 would have an API growth target of 5 points (5% of 800 – 700) for 2004, while a schools with APIs of 650 and 750 would have growth targets of 6 and 4, respectively. California's focus on growth rather than status obviously stands in sharp contrast to the federal AYP requirements.

In order for a school to meet all API growth requirements, students in the school who are members of a "numerically significant" subgroup defined by ethnicity or socioeconomic disadvantage "must achieve at least 80 percent of the schoolwide annual growth target (California Department of Education, 2004, p. 30). Thus, while the California accountability system includes some attention to subgroup performance within a school, it does so in a way that differs from NCLB in at least two significant ways. First, fewer subgroups are considered. Second, the subgroup target for improvement of performance is somewhat lower than the schoolwide target, which, unlike NCLB, implicitly makes some allowance for the less reliable gains in achievement for subgroups that obviously have fewer students than are available for calculating the schoolwide changes to the API.

The one provision of the NCLB accountability system that considers improvement from one year to the next rather than only annual performance in comparison to an AYP is the, so called, safe harbor provision. If a subgroup of students in a school falls short of the AYP target, the school can still meet AYP if (a) the percentage of students who score below the proficient level is decreased by 10% from the year before, and (b) there is improvement for that subgroup on other indicators.

Although the safe harbor provision is intended to allow schools that fall short of the AYP goal to still make AYP if they show substantial improvement, very few schools that would not otherwise make AYP do so because of the safe harbor provision. The very small percentage of schools that are saved by the safe harbor provision is due to the fact that the 10% decrease in students scoring below proficient sets a very high bar in comparison to what is achieved even by schools where students show considerable improvement from one year to the next. Only a tiny fraction of schools actually meet AYP through the safe harbor provision because it is so extreme.

If a provision is desired to allow schools to meet AYP by showing decreases in the percentage of students scoring below the proficient level, then consideration should be given to alternative criteria such as an above average decrease in the percentage of students scoring below the proficient level from one year to the next. This would likely lead to a criterion closer to a 3% reduction in the below proficient category from one year to the next rather than the current 10% criterion. Changing the safe harbor provision from a 10% reduction in below proficient to a 3% 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.

Achievement Goals

There are several important differences between the way in which school achievement goals are set for purposes of NCLB and the ways in which they are typically set in state accountability systems. First, as was discussed in the previous section, there is the difference between status in comparison to a target used for NCLB and improvement targets used in the typical state accountability system. Second, there is the difference between NCLB's use of an absolute level of performance that is constant regardless of a school's initial status and the targets set by states that are usually set at levels that depend on the school's performance in a baseline year or biennium. Third, NCLB and the typical state system differ in the establishment of long-range goals and the timeline for reaching those goals.

As was previously noted, NCLB requires states to set challenging academic performance standards. There must be at least three performance standards for each assessment. NCLB provides only general guidelines to states for defining the academic achievement standards, specifying only that a state establish "challenging academic achievement standards that – (I) are aligned with the state's academic content standards; (II) describe two levels of high achievement (proficient and advanced) that determine how well children are mastering the material in the state academic content standards; and (III) describe a third level of achievement (basic) to provide complete information about the progress of the lower-achieving children toward mastering the proficient and advanced levels of achievement" (NCLB, P.L. 107-110, sec. 1111 (b)(1)(D)).

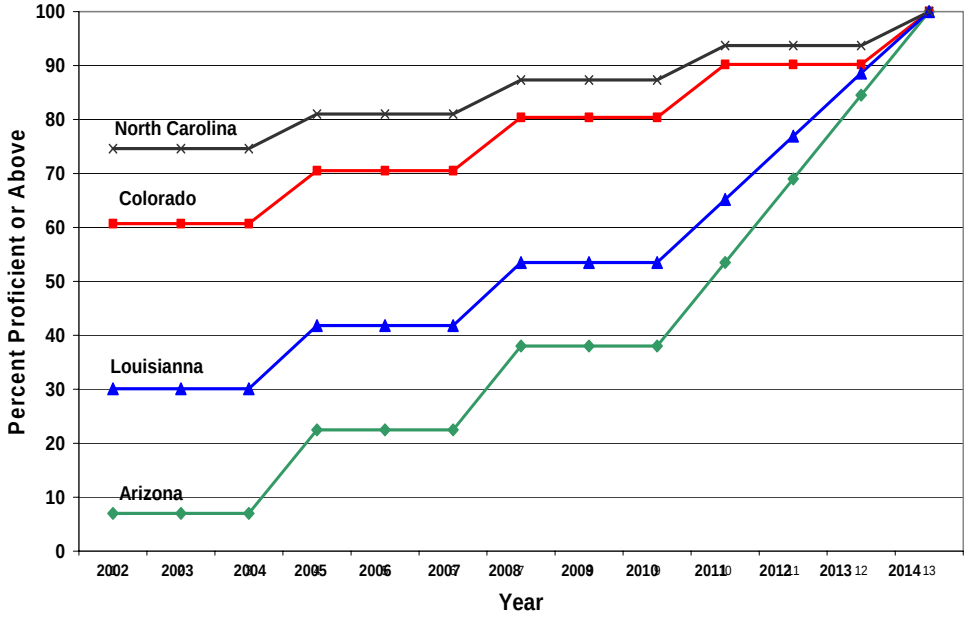
All students must be at the proficient level or above by school year 2013-2014 for schools and districts to avoid sanctions, regardless of how leniently or stringently the state defines the proficient standard. State starting points for percent proficient were supposed to have been established using assessment results from the 2001-2002 academic year. The state's starting point is equal to the higher of the following two values: (a) the percentage of students in the lowest scoring subgroup who achieve at the proficient level or above and (b) "the school at the 20th percentile in the state, based on enrollment, among all schools ranked by the percentage of students at the proficient level" (NCLB, P.L. 107-110, Sec. 1111 (b)(2)(E)(ii)).

States must also establish intermediate goals, called AMOs, for AYP between the 2001-2002 starting point and the 100% proficient goal in 2013-2014. The first increase in the goal from the starting point must occur by 2004-2005, and subsequent increases must occur in not more than 3 years following the last increase. Although some states (e.g., Florida) have set their intermediate goals for AYP by using equal increments each year to move from the starting point in 2002 to 100% in 2014, a number of states have elected to use a stair-step approach to setting their intermediate goals for AYP with increases in 2005, 2008, 2011, and 2014 and static levels for intermediate years as illustrated in Figure 3 by Colorado and North Carolina. Alternatively some states opted for stair steps in 2005, 2008, and 2011, but then had annual increments through 2014 as is illustrated in Figure 3 by Arizona and Louisiana.

A comparison of the graphs of the AYP targets for the four states displayed in Figure 3 shows that that the four states have quite different starting points. North Carolina has a start point of 74.6% proficient or above which is slightly more than 10 times as high as Arizona's starting point of 7% proficient. Colorado's starting point of 60.7% is twice as high as Louisiana's starting point of 30.1%. Yet the 2014 AYP target for all students in these four states, as well as in all the other states, is 100% proficient or above. That sort of improvement in student achievement is completely unrealistic (see, for example, Linn, 2003, 2004; McCombs, Kirby, Barney, Darilek, & Magee, 2004, for discussions of the unrealistic nature of the 100% proficient goal by 2014).

The differences among starting levels for the four states are much larger than the differences in actual performance of students in eighth-grade mathematics. The percentages of public school students who were at or above the proficient level on the 2003 NAEP mathematics assessment were as follows for the four states displayed in Figure 3: Arizona, 21%; Colorado, 34%; Louisiana, 17%; and North Carolina, 32% (National Center for Education Statistics., 2003). Although student performance on the 2003 Grade 8 NAEP mathematics assessment does differ for these four states, the range from high to low in percent at or above proficient is 17%, which is small in comparison to the range in differences in AYP starting points of 67.6%. Moreover, the rank order of the four states in terms of AYP starting values does not match the rank order in terms of actual performance on NAEP in 2003.

Figure 3
Intermediate AYP Goals for Four Illustrative States



It should also be noted that the steep annual increases that Arizona and Louisiana chose to use to set AYP targets for the last four years (2011 through 2014) are just the opposite of what reasonably might be expected. Past experience with test-based accountability systems has shown that larger gains are usually made in the first few years following implantation and that gains generally become smaller in later years. Moreover, common sense suggests that it likely to be much harder to realize a gain of 5% to move from 95% to 100% than from 10% to 15% proficient or above.

State-to-State Variability in Percentage of Schools Meeting AYP Goals

Figure 3 illustrates the fact that states differ substantially in their starting points, as well as their intermediate AYP goals. There is also considerable state-to-state variability in the percentage of schools that met AYP in each of the first 2 years (2003 and 2004) of AYP reporting. Olson (2004) reported the percentage of schools making AYP for 41 states in 2003 and 44 states in 2004. The percentage of schools that met AYP goals in 2004 in a 45th state, Illinois, was obtained from the Illinois State Board of Education website. In 2003 the percentage of schools that met AYP goals ranged from a low of 18% to a high of 95%, with an average of 65.6% for the 41 states

listed by Olson. For the 45 states in 2004, the range was from 23% to 96% with an average of 74.2%.

Olson (2004) reported the percentage of schools meeting AYP goals for **both** 2003 and 2004 for only 36 of the states. The percentage of schools meeting AYP goals was higher in 33 of the 36 states in 2004 than in 2003, and the three exceptions, Indiana, Louisiana, and Michigan, had decreases from 2003 to 2004 of only 1% or 2%. The average 2003 to 2004 increase in percentage of schools making AYP goals was 10.2% and eight states (Connecticut, Delaware, Massachusetts, Missouri, North Carolina, Pennsylvania, South Carolina, and Tennessee) had increases that ranged from 20% to 30%. These appear to be remarkable improvements in a single year. Although the increases reflect some real improvement in student performance, they are largely due to artifacts such as schools getting better about meeting requirements that at least 95% of the eligible students in each subgroup participate in the assessments. An even larger part of the apparent improvement can more reasonably be attributed to changes in AYP calculations that states requested and the U.S. Department of Education approved between the 2003 and 2004 reports.

The Center for Education Policy (CEP) posted a report on its website dated October 22, 2004, summarizing changes in state implementation of NCLB accountability rules (CEP, 2004). According to the CEP report, 47 states requested approval for change in their NCLB accountability plans, and the U.S. Department of Education posted letters to 35 of those states in time to be reviewed for the CEP report "approving many, though not all of the changes" (CEP, 2004, p. 1). Many of the approved changes make it easier for schools to meet AYP goals. For example, 11 states changed the minimum group size for disaggregated reporting, and 12 states introduced the use of confidence intervals.[2]

Either increasing the minimum group size or introducing the use of confidence intervals helped schools make AYP goals in 2004 that would not have made it without these changes in NCLB accountability plans. It is worth noting in this regard, that 4 of the states (Missouri, North Carolina, Pennsylvania, and South Carolina) that were among the 8 states showing the largest increases in percentage of schools meeting AYP goals from 2003 to 2004 were also among the 12 states that started using confidence intervals in 2004.

[2]A confidence interval gives the benefit of the doubt to schools in cases where the percentage of students who are proficient or above is somewhat below the target value required for meeting the AYP goal.

Meeting AYP Goals and Performance on NAEP

The substantial state-to-state variability in the percentage of schools meeting AYP makes it evident that the likelihood that a school will fail to meet AYP goals depends not only on the performance of students in the school, but also, at least in part, on the state in which the school is located. Furthermore, as can be seen in Figures 4 and 5, the percentage of schools in a state that meet AYP goals has a weak relationship to differences among states in student performance on NAEP. Figure 4 shows the relationship of the average percent proficient or above across the 2003 NAEP Grade 4 reading, Grade 8 reading, Grade 4 mathematics, and Grade 8 mathematics assessments and the percentage of schools within the state that met AYP goals in 2003.

Figure 4
Relationship of Average of State Percent Proficient or Above on 2003 NAEP and Percentage of Schools Meeting AYP in 2003
(41 States, Correlation = .26)

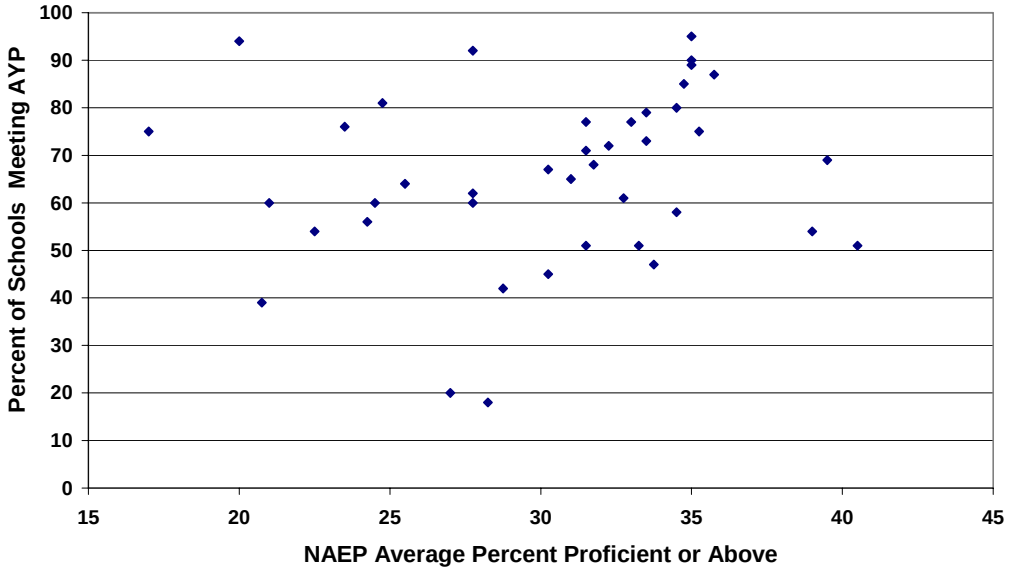


Figure 5
Relationship of State Average Percent Proficient or Above on 2003 NAEP and Percentage of Schools Meeting AYP in 2004 (45 States, Correlation = .35)

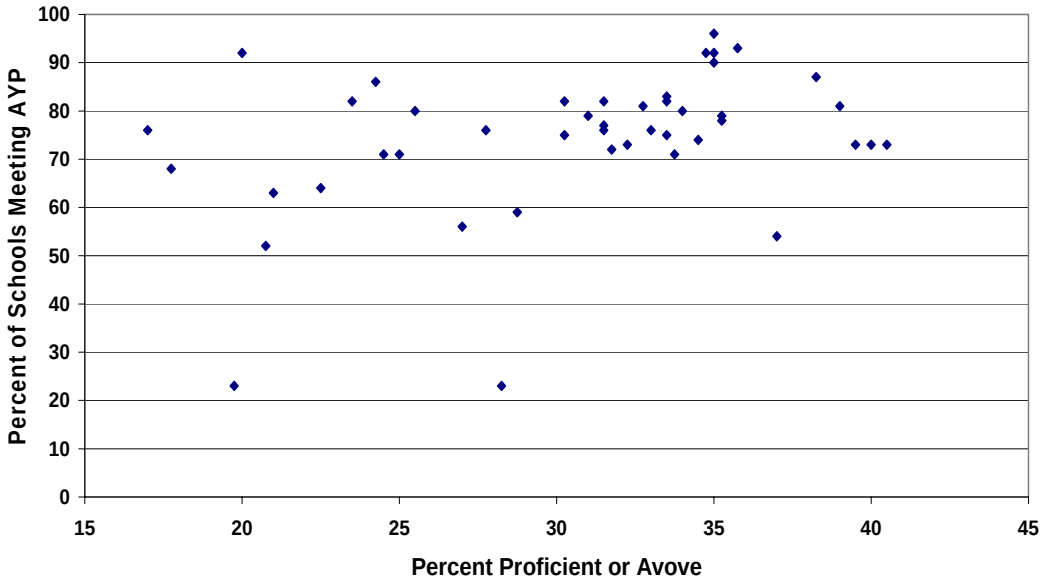


Figure 5 shows the relationship of the same NAEP average percent proficient or above with the percentage of schools that met AYP goals in 2004.

It is apparent from an inspection of Figures 4 and 5 that there is a relatively weak relationship between the performance of students in a state on NAEP and the percentage of schools that meet their AYP goals in the same state. In 2003, the state with the second lowest average performance on NAEP had more than 90% of its schools make AYP, whereas only half the schools met their AYP goals in the state with the highest performance on NAEP (Figure 4).

The relationship between the average NAEP performance in 2003 and the percentage making AYP in 2004 included a few more states and was slightly higher than it was for 2003 AYP results. Nonetheless, there are some notable outliers shown in Figure 5 where a state with relatively low average performance on NAEP has a noticeably higher percentage of schools making AYP than a state with relatively high average performance on NAEP.

Results such as those shown in Figures 3, 4, and 5, clearly illustrate that it is not meaningful to compare states in terms of their performance standards or the rates at

which schools in different states meet NCLB's AYP requirements. The performance standards set by states bear little relationship to real between-state variability in student performance. Differences in performance standards set by state as well as differences in the ways in which states, with the approval of the U.S. Department of Education, comply with NCLB accountability requirements, with regard to features such as the minimum number of students needed for determining if subgroups must meet targets, and whether or not confidence intervals are used, obscure between-state comparisons of percentages of schools meeting AYP requirements.

Conclusion

Test-based accountability has become a pervasive consideration for schools and educators as a consequence of the combination of state accountability requirements and those imposed by NCLB. Because of the substantial differences in state and NCLB requirements mixed messages that are confusing to the public are being given about school performance. The goals established under NCLB are already unrealistic for many schools that started with low performance in 2002 and will become increasing so, not only for those schools but for all schools as the increases in AYP targets start kicking in, especially in 2005 and 2008 when many states will have big jumps in their AYP targets. If the goal for 2013-2014 remains unchanged, essentially all schools will fail to meet the unrealistic goal of 100% proficient or above, and No Child Left Behind will have turned into No School Succeeding.

Significant changes in NCLB accountability requirements are needed to avoid labeling all schools as failures. What are some of the needed changes? Possibly most important is to make the goal something that is more realistically obtainable. NCLB requires states to participate every other year in the National Assessment of Educational Progress (NAEP) reading and mathematics assessments at Grades 4 and 8 starting in 2003. Although the use of state-level NAEP results are not specified in the law, it is reasonable to think of those results as providing some kind of benchmark for state assessments. In 2003, no state or large district had anything close to 100% of their students performing at the *basic* level, much less the *proficient* level at either Grade 4 or Grade 8 in either reading or mathematics.[3]

[3]It should be noted that the NAEP achievement levels have been the subject of considerable criticism, in part, because they are set at levels that are higher than the performance of students in any country (see, for example, Linn, 2003).

Performance goals "mandated by the accountability system should be ambitious, but also should be realistically obtainable with sufficient effort" (Linn, 2003, p.4). At the very least, there needs to be an existence proof. That is, there should be evidence that the goal does not exceed that that has previously been achieved by the highest performing schools. For example, if the top 10% of schools in a state (in terms of sustained improvements in student achievement) had rates of improvement in the percentage of students achieving at the proficient or above level during the past 5 years that averaged 3% per year, then adequate yearly progress might be defined as a 3% increase in the percentage of students achieving at proficient or above each year. That would be a great challenge to the vast majority of schools, but might be a target that is within reach with sufficient effort.

Saying that all students must be at the proficient level or above by 2014, but leaving the definition of proficient achievement to the states has resulted in so much state-to-state variability in the level of achievement required to meet the proficient standard that "proficient" has become a meaningless designation. Certainly, reporting results in terms of percent proficient on state assessments lacks comparability from state to state.

If the percentage of students who are above a cut score on a state assessment is to be used, the cut score should be more meaningful than the state-established proficient levels that lack any semblance of a common meaning across states. There are several approaches that would be preferable to reporting results in terms of percent proficient or above. One simple approach would be to define the standard or cut score on a state assessment to be equal to the median score in a base year, presumably 2002. The percentage of students scoring above that constant cut score would then be used to monitor improvement in achievement with target increases set at reasonable levels (e.g., 3% per year). With a target increase of 3% a year, the proportion of students scoring above the 2002 median would need to increase from 50% in 2002 to 86% in 2014. That would represent a gigantic improvement in the achievement of the nation's students, but might not be totally unrealistic, and surely is not as poorly defined as 100% proficient or above given the huge state-to-state variability in the meaning of proficient.

Another alternative would be to use what Popham (2004) has called grade-level descriptions. At grade level might correspond more closely to the "basic" than the "proficient" level in most states. Using past experience, targets could be set that

would bring the achievement of an ever-increasing percentage of students up to the "at-grade-level" standard.

The NCLB insistence on a common target for all schools, regardless of where they started is appealing in the sense that it sets the same high expectations for all, but is nonetheless counterproductive when it leaves schools with initially low performing students with no realistic hope of making the absolute target. Schools demonstrating substantial improvement should not be labeled as failing to make adequate progress, and for the reasons discussed above, NCLB's safe harbor provision turns out to be no real help to most schools in this regard due to the high hurdle it establishes.

Holding schools accountable for the performance of students in subgroups that have too often been ignored in the past (e.g., racial/ethnic minorities, economically disadvantaged, limited English proficient students, and students with disabilities) is a desirable feature of NCLB. However, as it is implemented, it places large, diverse schools at a substantial disadvantage. Changing the safe-harbor provision from a 10% reduction in below proficient to, say, a 3% 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.

References

Anderson, B. (1982). Test use today in elementary and secondary schools. In A. K. Wigdor & W. R. Gardner (Eds.), *Ability testing: Uses, consequences, and controversies. Part II: Documentation section* (pp. 232-285). Washington, DC: National Academy Press.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508- 600). Washington, DC: American Council on Education.

California Department of Education. (2004, October). *2003-2004 Academic performance growth report: Information guide*. Available at http://www.cde.ca.gov/

Center on Education Policy. (2004). *Rule changes could help more schools meet test score targets for the No Child Left Behind Act.* Retrieved October 22, 2005, from http://www.cep-dc.org/nclb/StateAccountabilityPlanAmendmentsReportOct2004.pdf

Clarke, M., Madaus, G., Horn, C., & Ramos, M. (2001, April). *The marketplace for educational testing*. National Board on Educational Testing and Public Policy Statements, Vol. 2, No. 3., Carolyn A. and Peter S. Lynch School of Education, Boston College,. Available from http://www.bc.edu/research/nbetpp/reports.html

Dillon, S. (2004, September 5). Good schools or bad? Conflicting ratings leave parents baffled. *The New York Times.*

Ford, L. A., & Thacker, A. A. (2005). *Consequential impact of the No Child Left Behind Act on Kentucky's Accountability System: Phase I*. FR-05-03. Louisville, KY: HumRRO.

Hemenway, J. S., Wang, M., Kenoyer, C. E., Hoepfner, R., Bear, M. B., & Smith, C. (1978). *Report #9: The measures and variables in the Sustaining Effects Study.* Santa Monica, CA: Systems Development Corp.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 475-514). New York: Macmillan.

Kentucky Department of Education. (2004). *2004 CATS interpretive guide: Detailed information on using your score reports.* PDF document retrieved April 19, 2005, from

http://www.education.ky.gov/NR/rdonlyres/estue3du34uop7gly654o4qk254zn2s7hrsg4d6mow2sdepzmrsmyvqes44pl7mbch5n2ha77mqgrl77vksjoemncze/InterpretiveGuide2004.pdf

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

Linn, R. L. (2004, July 28). *Rethinking the No Child Left Behind accountability system.* Paper presented at Center for Education Policy Forum, Washington, DC. Paper available at http://www.cep-dc.org/pubs/Forum28July2004/BobLinnPaper.pdf

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

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and post-testing periods. *Review of Educational Research*, *47*, 121-150.

Marion, S. T., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, A., & Sheinker. J. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. A paper series: Implementing the state accountability requirements under the No Child Left Behind Act of 2001. Washington, DC: Council of Chief State School Officers.

Mazzeo, C. (2001). Frameworks of state assessment policy in historical perspective. *Teachers College Record*, *103*, 367-397.

McCombs, J. S., Kirby, S. N., Barney, H., Darilek, H., & Magee, S. (2004). *Achieving state and national literacy goals, a long uphill road*. TR-180. Santa Monica, CA: RAND.

National Center for Education Statistics. (2003). The Nation's Report Card, 2003 (math). Washington, DC: Author. Retrieved April, 19, 2005, from http://nces.ed.gov/nationsreportcard/mathematics/results2003/stateachieve-g8-compare.asp

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Novak, J. R., & Fuller, B. (2003, December). *Penalizing diverse schools? Similar test scores, but different students, bring federal sanctions* (Policy Brief No. 03-4). Berkeley: University of California. Retrieved April 19, 2005, from http://pace.berkeley.edu/policy_brief_03-4_Pen.Div.pdf

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

Olson, L. (2004). Taking root (special report). *Education Week*, *24*(15), S1-S10.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: Macmillan.

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. *Phi Delta Kappan*, *59*, 585-588.

Pipho, C. (1984). *State activity: Minimum competency testing* (unpublished table). Denver, CO: Education Commission of the States.

Popham, W. J. (2004, May 26). Shaping up the "no child" act: Is edge-softening enough? *Education Week*, *23*(38), 40.

Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Gardner (Eds.), *Ability testing: Uses, consequences, and controversies. Part II: Documentation section* (pp. 173-194). Washington, DC: National Academy Press.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-265). Philadelphia: Falmer Press.

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