Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001

CSE Technical Report 567

Robert L. Linn
CRESST / University of Colorado at Boulder
Eva L. Baker
CRESST / University of California, Los Angeles
Damian W. Betebenner
University of Colorado at Boulder

June 2002

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

Project 1.2: Systems Design and Improvement: Ideal and Practical Models for Accountability and Assessment
Eva L. Baker, CRESST/UCLA, and Robert L. Linn, CRESST/University of Colorado at Boulder, Project Directors

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

ACCOUNTABILITY SYSTEMS: IMPLICATIONS OF REQUIREMENTS OF THE NO CHILD LEFT BEHIND ACT OF 2001

Robert L. Linn
CRESST/University of Colorado at Boulder

Eva L. Baker
CRESST/University of California, Los Angeles

Damian W. Betebenner
University of Colorado at Boulder

Abstract

The recently enacted No Child Left Behind Act of 2001 amends the Elementary and Secondary Education Act of 1965. The new law substantially increases the testing requirements for states and sets demanding accountability standards for schools, districts, and states, including the setting of measurable adequate yearly progress objectives for all students, as well as for subgroups of students defined by socioeconomic background, race/ethnicity, and English language proficiency. Some of the implications of the law for state accountability systems are discussed. Issues raised by variations among states in their content standards, the rigor of their tests, and the stringency of their performance standards are illustrated. In addition, the differences in types of tests are considered, as well as provisions related to how local tests might be integrated into the system. Some suggestions are provided for leveling the playing field among states and for improving the ways in which adequate yearly progress is evaluated.

By making accountability the centerpiece of the education agenda, President George W. Bush (The White House, 2001) strongly 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 have now become law with the signing of the No Child Left Behind Act of 2001 (NCLB) in January 2002 (Public Law 107-110). NCLB amends the Elementary and Secondary Education Act of 1965. It has a number of testing and accountability provisions that will require changes in the practices of many states. The law requires, for example, that states develop content standards in reading and mathematics and tests that are linked to those standards in Grades 3 through 8. Science content standards and assessments will follow.

Most states have already developed content standards in reading and mathematics, as well as in some other subjects, and have tests that are arguably linked to those standards. Many states, however, do not administer tests in both reading and mathematics each year to students in Grades 3 through 8. Indeed, according to a recent summary in *Education Week*, only nine states currently have standards-based tests in both English and mathematics at Grades 3 through 8 (Olson, 2002). By the time the NCLB requirements are fully in effect, in the 2005-2006 academic year, states that currently test only in selected grades will have to have completed the necessary development to administer tests in both reading and mathematics to all students in Grades 3 through 8.

Despite the focus during the legislative process on the key provisions related to adequate yearly progress and the challenges such targets present both methodologically and practically, let us also briefly address differences in assessment approaches that are used by various states.

The goal of assessment is to provide a valid set of inferences related to particular expectations for students and schools. The way states expect such assessments to map to standards varies. In addition to difficulty levels (associated with the actual items and tasks used on an assessment) and the stringency of performance standards, testing programs vary, at least nominally, on the strategies they use to measure performance. Putting aside discussions of open-ended (constructed) versus multiple-choice (selected) response modes, with the proposition that both can be used to measure challenging or trivial educational accomplishments, there are still potential differences that are important. One is whether the assessment system is domain focused and standards based in design (i.e., the items are specially constructed to relate to clearly specified outcomes) or whether standards are used as a strategy for reporting. The difference in strategy relates not only to differences in theory about how measurement should occur but also to how sensitive instruction is likely to be in prompting changes in performance. Both positions have strong proponents. The reality may be that tests labeled norm referenced or criterion referenced may share a common item pool and thus perform comparably. Certainly an improved understanding of the expectations for various measures has implications for how the process of “alignment” is regarded, as well as expectations for change. Therefore, the discussion of performance standards, adequate yearly progress, and an external arbiter for performance (e.g., the National Assessment of Educational Progress) needs to be

considered in the light of very different, but scientifically supportable, measurement models.

States will also need to make a number of other changes in their testing and accountability systems as a result of the NCLB requirements. Notable among the other changes are those concerned with the identification of adequate yearly progress objectives, requirements for disaggregated reporting of results, and the requirement to participate every other year in state-level administrations of the National Assessment of Educational Progress (NAEP) in reading and mathematics at Grades 4 and 8.

Adequate Yearly Progress

NCLB specifies that states must develop "adequate yearly progress (AYP) objectives" consistent with the following stipulations in the law.

1. States must develop AYP statewide measurable objectives for improved achievement by all students and for specific groups: economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency.

2. The objectives must be set with the goal of having all students at the "proficient" level or above within 12 years (i.e., by the end of the 2013-2014 school year).

3. AYP must be based primarily on state assessments, but must also include one additional academic indicator.

4. The AYP objectives must be assessed at the school level. At the end of 2 years, schools that have failed to meet their AYP objective for 2 consecutive years will be identified for improvement.

5. School AYP results must be reported separately for each group of students identified above so that it can be determined whether each student group met the AYP objective.

6. At least 95% of each group must participate in state assessments.

7. States may aggregate up to 3 years of data in making AYP determinations.

Performance Standards

The second stipulation—all students performing at the "proficient" level or higher within 12 years—requires the establishment of performance standards for

state tests. Although many states have already established performance standards for their tests, the standards were not set with an awareness that they would be used to determine AYP objectives with the stipulation that all students reach the proficient level or higher by 2014. In a number of cases, the proficient level has been set so high that it may be completely unrealistic to expect all students to reach that level by 2014. Certainly, it is the case that no state, or country, for that matter, is close to meeting that goal now (Linn, 2000). Indeed, only a very few schools with student bodies selected on the basis of past achievement, or schools that serve students from privileged backgrounds, now have all their students scoring at the elevated levels defined as "proficient" on the more rigorous state tests with demanding performance standards.

The content standards used by states to develop tests vary in specificity and in rigor. Content standards and associated tests are much more ambitious in some states than in others. The performance standards states have set that determine the cut scores used to define proficient performance on the test also vary widely from one state to another. The combination of these differences among states regarding their content standards, the rigor of their tests, and the levels of performance required for a student to be considered proficient means that states are not starting on a level playing field. If current tests and standards are used to set AYP objectives, some states will have much farther to go and will have to set much more demanding AYP objectives than others, not necessarily because their students are achieving less, but because of the greater stringency of their definitions of proficient performance.

Figures 1 and 2 display the trends in the percentage of students meeting state standards on state tests in Grade 8 reading and mathematics, respectively, for five states over a 4-year period from 1998 through 2001.[1] As can be seen, though there is some variation in the slopes of the trend lines from state to state over the 4 years, the main distinguishing characteristic of the trend lines is their level. In 2001 the percentage of students meeting standard on the state Grade 8 reading tests ranged from a low of 27 to a high of 91. The corresponding range for the Grade 8

[1] California uses the Stanford Achievement Test, 9th ed. (SAT9). The 50th national percentile rank was used as the cut point for the graphs. Maryland uses the Maryland School Performance Assessment Program (MSPAP). The percent of students scoring satisfactory or better is plotted in the graphs. Massachusetts uses the Massachusetts Comprehensive Assessment System (MCAS). The percent of students scoring proficient or better is plotted in the graphs. Oregon uses the Oregon Statewide Assessment. The percent of students meeting or exceeding performance standards is plotted in the graphs. Texas uses the Texas Assessment of Academic Skills (TAAS). The percent of students meeting the minimum expectations is plotted in the graphs.

mathematics tests was from 31 to 92. A straight-line projection of gains needed between 2001 and 2012 would require an annual increase of slightly less than 1% per year for the state with the highest percentages meeting standards in 2001 to nearly 5% per year for the state with the lowest percentages meeting standards in 2001.

Two of the states (Maryland and Texas) with results shown in Figures 1 and 2 have had testing programs in place since at least 1994. Trends for those two states for the 8 years starting in 1994 and ending in 2001 are shown for Grade 8 reading and mathematics in Figures 3 and 4, respectively. Over the 8 years, the trend lines differ both in level and slope, with the trends for Texas starting higher and having steeper slopes than those for Maryland.

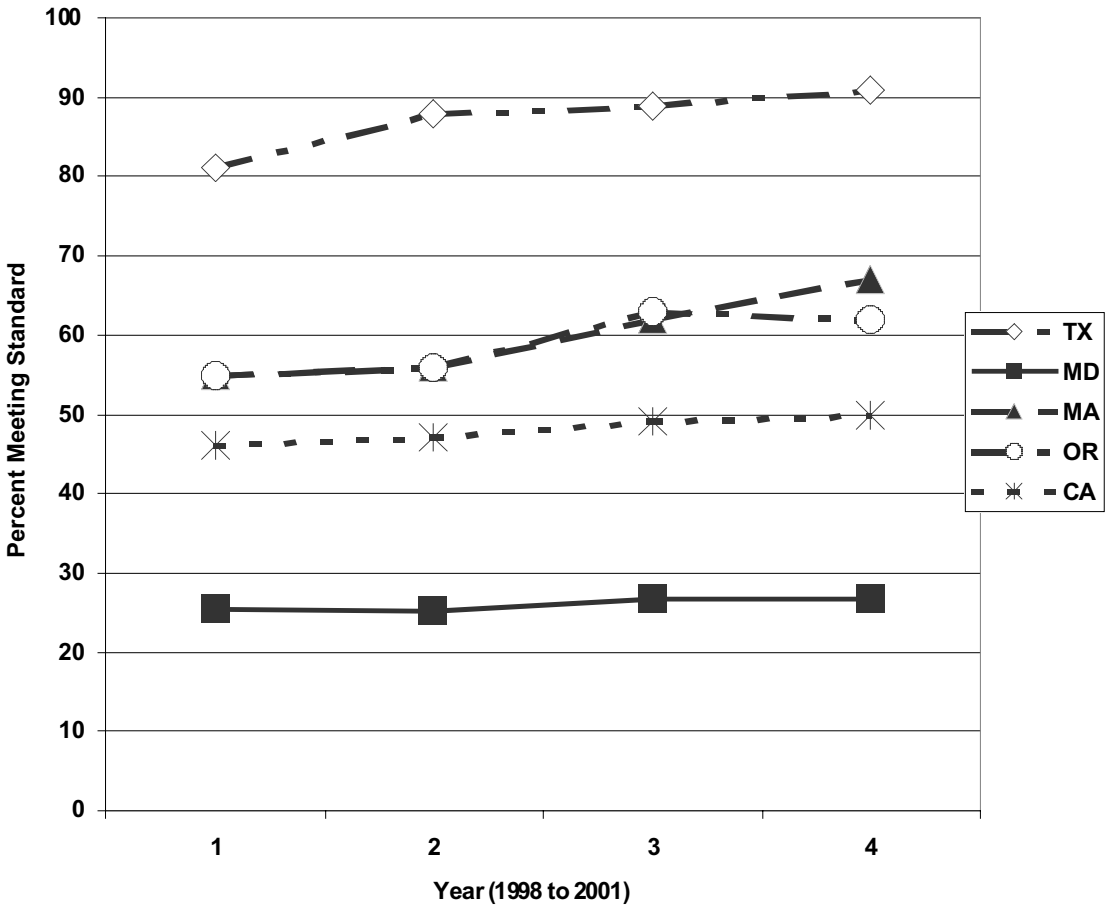

Figure 1. Trends in percent of students meeting standard in five states: Grade 8 reading.

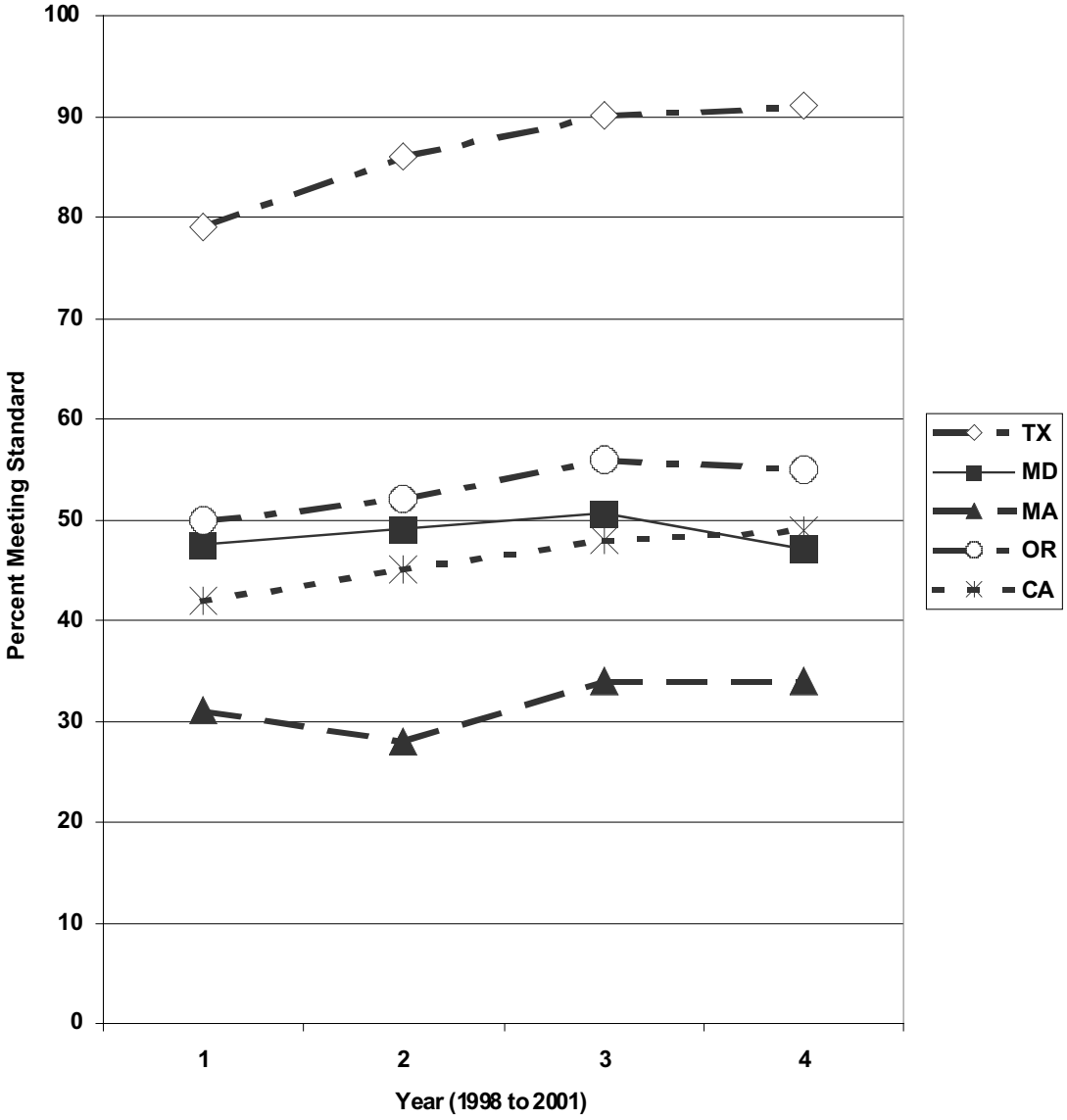

Figure 2. Trends in percent of students meeting standard in five states: Grade 8 mathematics.

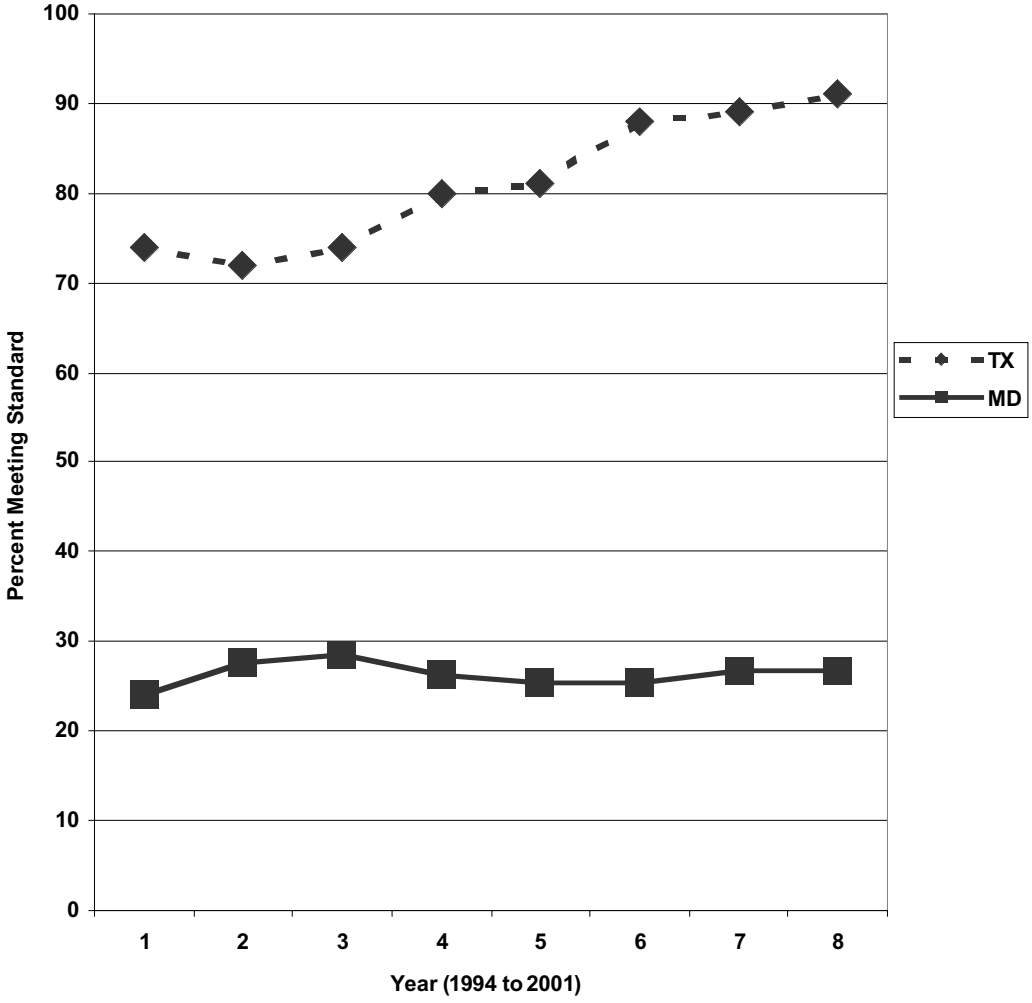

Figure 3. Trends in percent of students meeting standard in Texas and Maryland: Grade 8 reading.

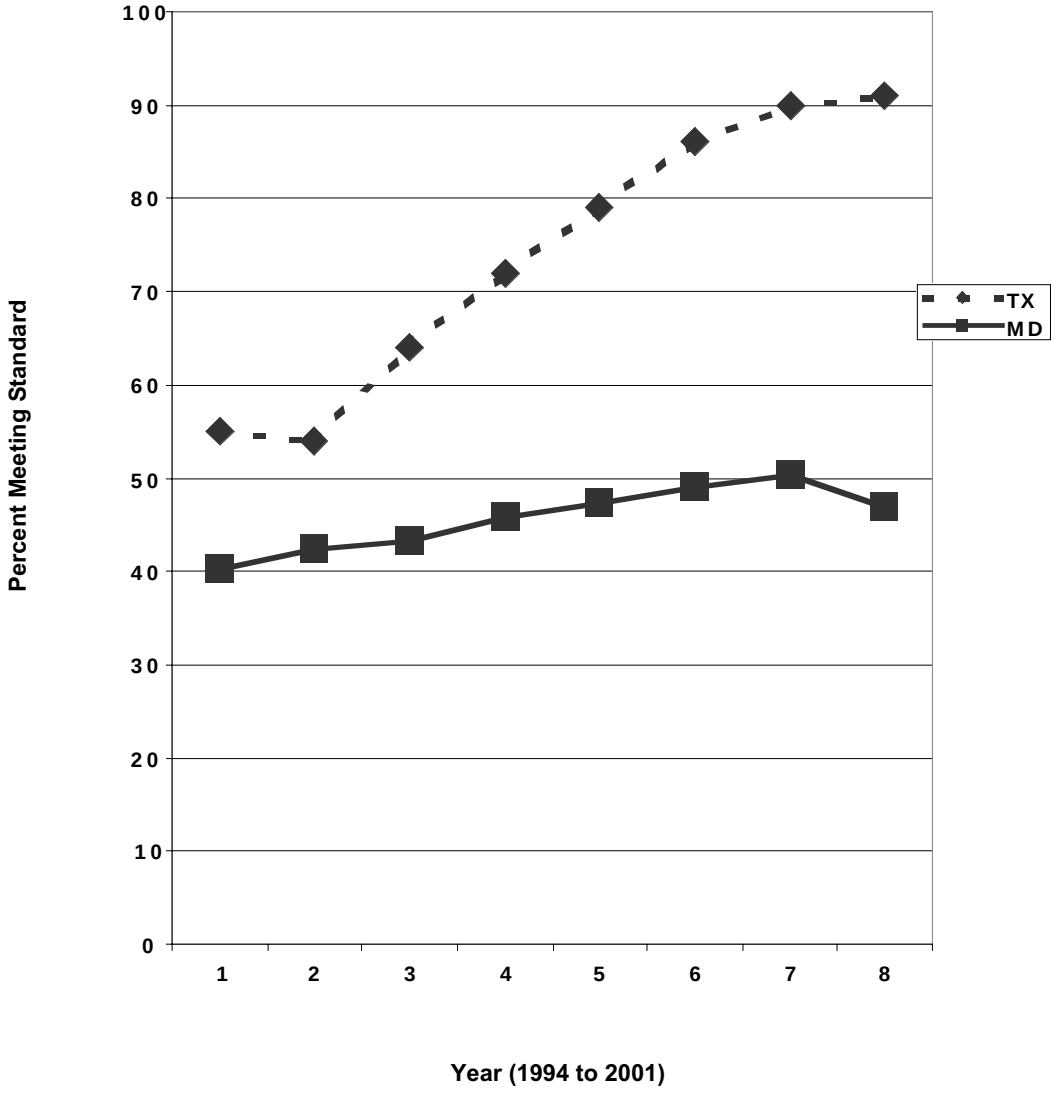

Figure 4. Trends in percent of students meeting standard in Texas and Maryland: Grade 8 mathematics.

State-level administrations of NAEP provide a basis for comparing the achievement of students from different states that is not clouded by the differences among the tests and performance levels adopted by different states for their own state assessments. Both Maryland and Texas participated in the Grade 8 state NAEP administrations in 1990, 1992, 1996, and 2000. NAEP reports results in several different ways, one of which uses NAEP achievement levels that divide student performance into four levels: below basic, basic, proficient, and advanced.

Figure 5 displays the trends in the percentage of students in Maryland and in Texas that scored at the proficient level or higher at each of the four administrations of the Grade 8 mathematics assessments at the state level. By comparing the trends in Figures 4 and 5, it is clear that the percentage proficient or higher on NAEP is less than the percentage meeting the standard on either of the state tests. For Texas, the slope of the trend line is also flatter on NAEP than on the state assessment. If the basic level or above is used rather than the proficient level or above on state NAEP, the percentages of students scoring in that range are obviously greater. Figure 6 displays the trends in percentage of students in Maryland and in Texas who scored at the basic level or higher on the Grade 8 NAEP mathematics assessments in 1990, 1992, 1996, and 2000. Comparing Figures 4 and 6, it can be seen that the slope of the trend line is considerably flatter for NAEP than for the state tests. It can also be seen that in recent years the percentages of students meeting standards on the Maryland test are lower than the percentage of Maryland students scoring at the basic level or higher, whereas the converse is true for Texas.

Although Texas is in the process of introducing new, more demanding tests, the tests that were in place during the years for which results are graphed in Figures 3 and 4 primarily measured basic skills. In contrast, NAEP and the Maryland tests used during the period covered in the figures were more challenging. It would appear that substantial gains may be more obtainable on a basic skills test than on a test that measures more complex reasoning and problem-solving skills.

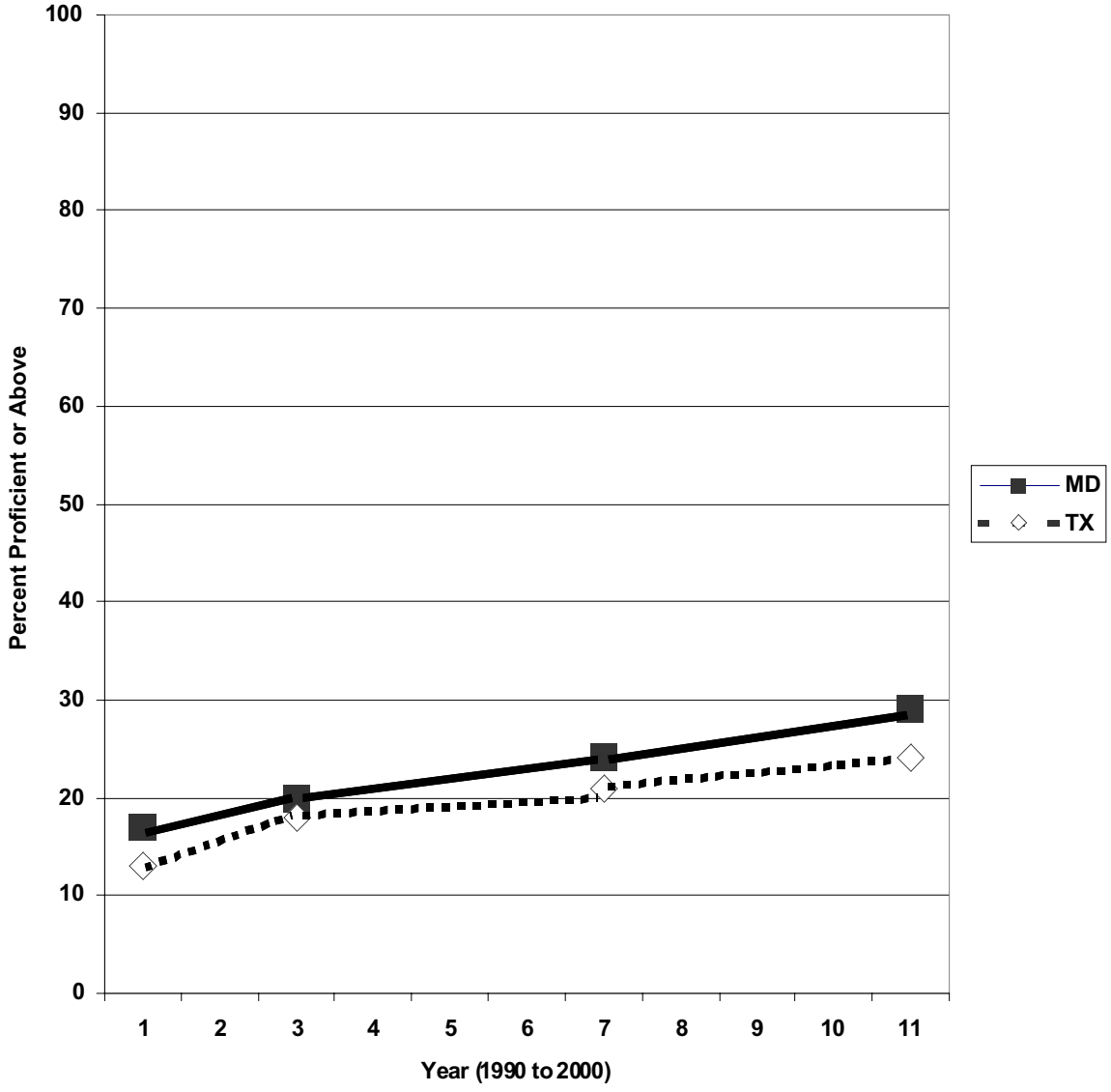

Figure 5. Trends in percent of students proficient or above on state NAEP Grade 8 mathematics for Maryland and Texas (1990 through 2000).

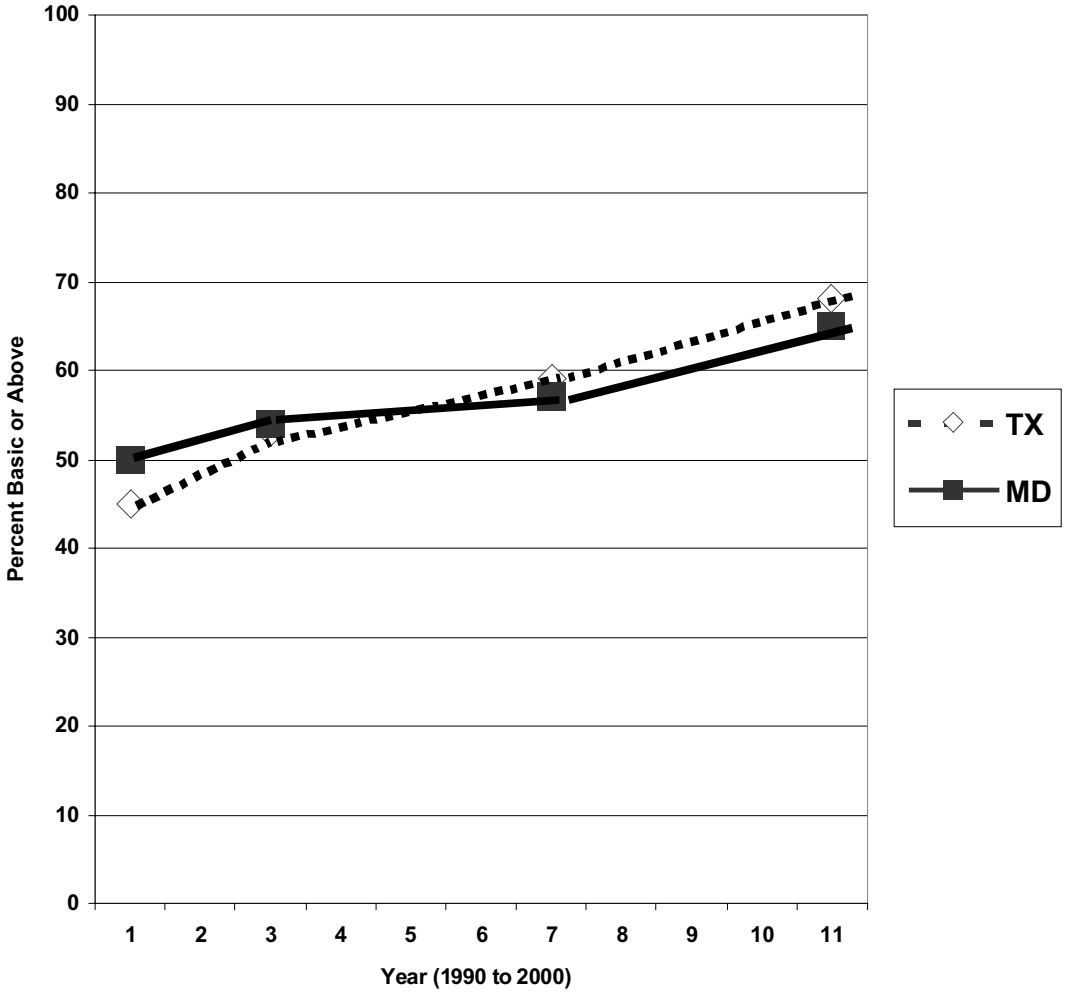

Figure 6. Trends in percent of students basic or above on state NAEP Grade 8 mathematics for Maryland and Texas (1990 through 2000).

Establishing AYP Objectives

Before it was agreed in the House and Senate Conference Committee to allow states to specify AYP objectives, the House and Senate versions of the bill for NCLB had set targets based on the idea that schools should increase the percentage of students scoring at the proficient level or higher by at least one point per year. It was also expected that schools would have to show an increase of at least one percentage point per year, not just for the overall student body, but also for all subgroups of students designated in the law for separate reporting (e.g., economically disadvantaged, African American, Hispanic, and White), and close the gap in achievement.

Even a steady increase of at least a point per year in the percentage of students proficient or higher each year would still leave states far short of the target of all students at that level by 2014, and these targets for schools are unlikely to be met by many schools. Using the criterion of a one-point-a-year increase, many schools would be identified for improvement. Indeed, the number of schools likely to be so identified is apt to be many times greater than the number that can be provided meaningful assistance even if the resources for school assistance programs were expanded substantially.

The changes in the percentage of students who score at the proficient level or higher on NAEP have differed by state, but have been relatively modest during the last decade for most states. Figure 7 displays the changes from 1992 to 1998 in the percentage of students scoring at the proficient level or higher for 33 states that participated in the state-level NAEP Grade 4 reading assessment in both 1992 and 1998. The heavy horizontal line shows the gain that would be needed for an average increase of 1% per year over the 6 years between assessments. As can be seen, only 3 of the 33 states had increases in the percentage of students scoring at the proficient level or higher that averaged one point or more a year.

As shown in Figure 8, states showed larger gains on the Grade 4 NAEP mathematics assessment between 1992 and 2000. Fifteen of the 34 states that participated in both the 1992 and 2000 Grade 4 mathematics assessments showed average yearly increases in the percentage of students scoring at the proficient level or higher that averaged one point or more per year. The Grade 8 NAEP mathematics assessments were administered at the state level in 1990, 1992, 1996, and again in 2000. Eighteen of the 29 states that participated in the Grade 8 mathematics

assessment in both 1990 and 2000 had increases in the percentage of students scoring at the proficient level or higher (see Figure 9). Judging from these NAEP results, it seems clear that an expectation that all schools will show increases of one point or more in both reading and mathematics is an extraordinarily ambitious goal. As has already been illustrated, it is not uncommon for the percentage of students scoring at currently identified levels for proficient or better on a state test to be 50% or 40%, or even less, for the state as a whole. For a state starting with only 40% or 50% of its students scoring at the proficient level or above, it is clear that steady progress at a one-point increase per year would mean many more than 12 years would be needed to reach the goal of all students achieving at least the proficient level.

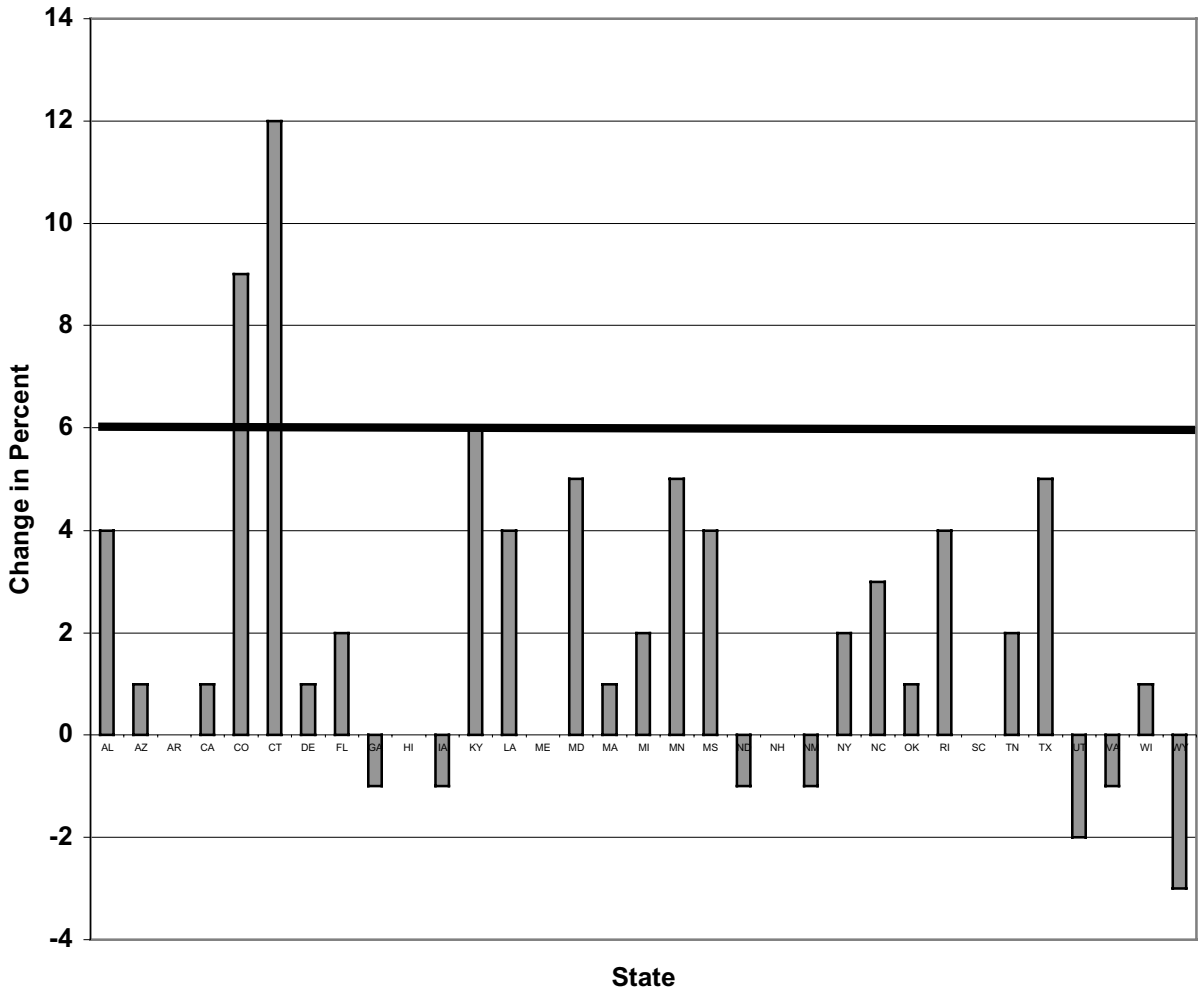

Figure 7. State changes in percent proficient or above from 1992 to 1998 on NAEP Grade 4 reading assessment.

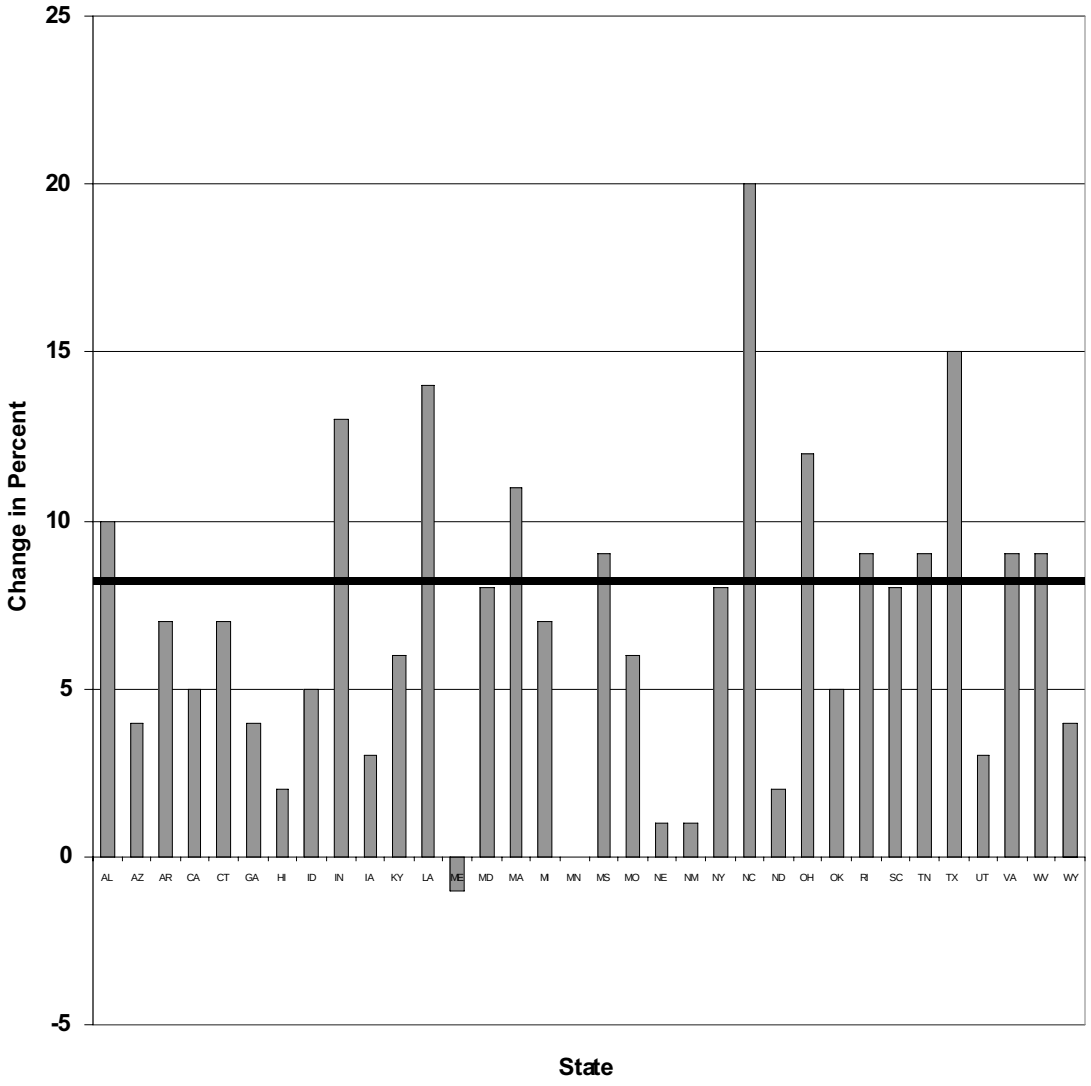

Figure 8. State changes in percent of students proficient or above from 1992 to 2000 on NAEP Grade 4 mathematics assessment.

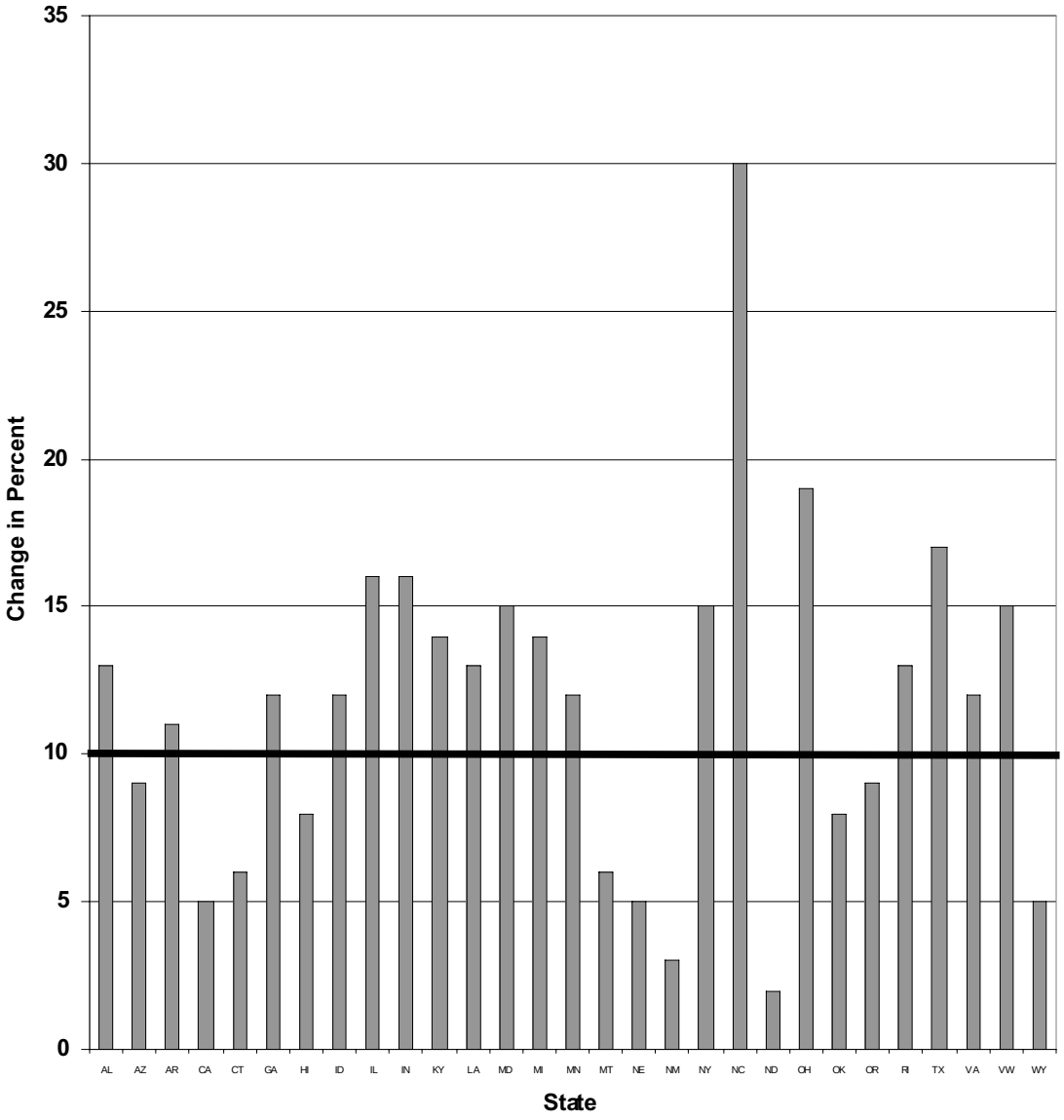

Figure 9. State changes in percent of students proficient or above from 1990 to 2000 on NAEP Grade 8 mathematics assessment.

One response to the gradual progress of the past is, of course, that schools have not been doing well enough in the past and must do better in the future. Indeed, that is a clear motivation behind not only the NCLB law, but also laws that have been passed in the last few years in quite a number of states. The notion is that, given enough pressure from the accountability system and some additional resources, the schools will improve and the goals will be met. One could agree that schools should improve and that holding schools accountable will contribute to improvement, but still conclude that the goal of having 100% of students reaching the proficient level or higher, as proficient is currently defined by NAEP or by many state tests, is so high that it is completely out of reach. Furthermore, setting a goal that is unobtainable, no matter how hard teachers try, can do more to demoralize than to motivate greater effort. Goals need to provide a challenge but not be set so high that they cannot possibly be achieved.

Individual School Results

AYP objectives at the school level present substantial challenges. There seems to be little recognition that school-level results are often volatile from year to year because of differences in cohorts of students from one year to the next. School teaching staff may also vary, so the inference that school A has made or not made progress across a 3-year period may apply to a relatively small proportion of students and teachers. Unfortunately, changes in scores for students tested at a given grade from one year to the next can be markedly unreliable. There are several sources for this unreliability. School summary scores for each year are subject not only to measurement error, but to sampling error. Sampling error is actually a much larger contributor to volatility of school-building scores than measurement error (Cronbach, Linn, Brennan, & Haertel, 1997).[2] In addition, difference scores that are computed as an indicator of progress tend to be less reliable than the scores used to compute the differences (e.g., Cronbach & Furby, 1970; Linn & Slinde, 1977). This result occurs because both the base-year test scores and the follow-up test scores are subject to errors of measurement.

[2] It is a surprise to some that sampling error is relevant since all, or nearly all, students in a tested grade in a school are tested. However, as Cronbach et al. (1997) have argued, for an assessment to be used as the basis for concluding "that a school is effective as an institution requires the assumption, implicit or explicit, that the positive outcome would appear with a student body other than the present one, drawn from the same population" (p. 393).

Moreover, the between-school variability of change scores is considerably smaller than the between-school variability of scores for a given year. School means for a given year vary greatly because of the large between-school differences in the socioeconomic backgrounds of the students who attend different schools. The mean scores of students who attend a school one year tend to be relatively similar to the mean scores of students who attended the previous year. Hence, the changes in mean scores from one year to the next are less variable than the means for either year. Finally, a substantial part of the variability found in change scores for schools is due to nonpersistent factors (e.g., turnover in the teaching staff, a teacher strike, or an especially disruptive cohort of students) that influence scores in one year but not the other (Kane & Staiger, 2001; Linn & Haug, 2002).

Results from the Colorado Student Assessment Program (CSAP) provide an illustration of the instability of school-building results. The CSAP has administered tests in reading at the fourth grade since 1997. The CSAP reading results for the 4 years from 1997 through 2000 provide a means of demonstrating the number of schools that would meet the one-percentage-point gain in proficient or better in just a single subject and without even requiring that all subgroups within the school meet that standard. Table 1 reports the number and percentage of schools that had an increase of one point or more in the percentage of Grade 4 students who scored at the proficient level or higher for three years.

As can be seen, slightly less than half of the schools met the target in 1998, whereas slightly more than half the schools met the target in 1999 and 2000. The results would look considerably worse if schools had to meet the target not only in reading but also in mathematics, and not only for the aggregate of all fourth-grade

Table 1

Number and Percentage of Schools That Met or Exceeded Target of an Increase in the Percentage of Students Scoring at the Proficient Level or Higher by Year (CSAP Fourth-Grade Reading Tests)

Years	Number of schools meeting target	Percentage of schools meeting target
1997 to 1998	333	44.8
1998 to 1999	431	56.5
1999 to 2000	431	55.5

Note. The total number of schools in the analyses was 744 for 1997-1998, 763 for 1998-1999, and 776 for 1999-2000.

students but for every subgroup of students in the school. Furthermore, as was previously indicated, steady progress at a one-point increase per year is not sufficient to bring the percentage of students reaching the proficient level or higher to 100% at the end of 12 years.

NCLB includes an expectation that schools should continue to meet their AYP objectives year after year. Many schools that meet the target in one year, however, will fail to do so the next year. This can be clearly seen in Table 2, where results for three successive years of meeting the target of at least a one-point increase in the percentage of students at the proficient level or higher on the fourth-grade CSAP test in reading are summarized for the 734 schools with results for all 4 years.

Even with a single test and without separate subgroup reporting, only 1 school in 20 would have met the target increase 3 years in a row. This is so despite the fact that on average, schools had 4.7% more students at the proficient level or higher in 2000 than they did in 1997. That is, the typical school had an average increase of more than 1.5% per year over the 3 years but failed to show gains of at least 1% in each of the 3 years.

Reducing the Volatility of School-Building Results

The fourth requirement—that school-level AYP results be available at the end of 2 years so that schools can be identified for improvement—has advantages over basing such identification on change in a single year. The volatility in school-building results from year to year is considerable (Kane & Staiger, 2001; Linn & Haug, 2002). Indeed, the volatility due to sampling error and nonpersistent factors is

Table 2

Number and Percentage of Schools Meeting or Exceeding Target of an Increase in the Percentage of Students Scoring at the Proficient Level or Higher in 0, 1, 2 or All 3 Years That Changes in Percentages Were Computed (CSAP Fourth-Grade Reading Test)

Number of years meeting target	Number of schools	Percentage of schools
0	21	2.9
1	315	42.9
2	362	49.3
3	36	4.9

so great that schools identified in a given year are unlikely to be similarly identified the following year. By accumulating 2 years of progress results for schools, the volatility will be reduced, though by no means eliminated. Stipulation 7, that states may aggregate up to 3 years of data in making AYP determinations, provides additional help in achieving dependable classifications because 3 years of data will lead to more trustworthy classifications of schools than only 2 years of data.

There are several alternative approaches to defining adequate yearly progress that could help ameliorate instability problems caused by differences in successive cohorts of students. Four possible alternatives that would likely help in this regard are (a) longitudinal tracking of students from year to year, (b) the use of rolling averages of 2 or more years of achievement results, (c) the use of composite scores across subject areas and grades, and (d) the use of separate grade-by-subject-area results but the setting of targets other than all combinations showing improvement (e.g., 5 out of 8, or 7 out of 10 possible grade-by-subject combinations). Each of these alternative approaches would reduce the magnitude of year-to-year fluctuations of results due to differences in cohorts of students attending a school.

Leveling the Playing Field

If stipulations in the NCLB law were taken at face value and current state tests and performance were used as starting points, it is clear that the requirements would vary greatly in stringency across states. It also is clear that the AYP objectives for states with reasonably ambitious tests and performance standards would not be feasible to meet. Hence, it is highly desirable that the negotiated rule making for the law provide interpretations and guidance for states that will both level the playing field across states and make it possible to define AYP objectives that are challenging but feasible to achieve given sufficient effort and concentration of resources. That is, the interpretations of the law need to enhance the likelihood for success of the underlying intent of the law to improve the achievement of all children and close the gap in achievement among racial/ethnic groups of students, and between children of poor parents and those of well-off parents. The interpretations of the law also need to minimize the likelihood of unintended negative consequences—for example, providing states with incentives to adopt less challenging content standards, develop tests aimed more at minimums than higher level understanding, and set cut scores at levels familiar in the era of minimum-competency testing.

The NCLB already requires biennial participation in state NAEP. The ways in which NAEP might be used to independently monitor state achievement trends or serve as a benchmark for comparing state tests or performance standards are not specified in the law. NAEP is the only common achievement measure that can serve as a benchmark for creating a more level playing field. In principle, there are many ways that NAEP might be used. At one extreme, for example, state NAEP results could be used to define the percentages of students at Grades 4 and 8 who achieve at various levels, and those results could be translated into the cut scores on the state tests that would yield equal proportions of students in the various achievement categories both at Grades 4 and 8, and by interpolation at Grades 5, 6, and 7, and by extrapolation at Grade 3.

Making NAEP the controlling factor would be fair to states in the sense that they would all be operating by the same rules. There are, however, a number of problems with such an approach. First, it in essence acts as if NAEP and the state tests were interchangeable, and, therefore, equitable. Unfortunately, as was concluded by a panel of the National Research Council that was charged with evaluating the feasibility of linking state tests to each other or to NAEP, there is far too much variability in the tests used by different states to justify an attempt to equate them to each other or to NAEP for purposes of reporting scores of individual students (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Although a less stringent standard might be appropriate to use for the linking of test results to report results for states or even schools, reporting of state test results under NCLB would not be limited to accountability reports for schools, districts and states, but would also include reports of individual student results to teachers and parents.

Relying on NAEP to determine performance levels and AYP objectives would also be problematic because of the stringency of the achievement levels (NAEP's performance standards). In the 2000 NAEP assessments, the percentage of students scoring at the proficient level or higher was 32% in reading at Grade 4, 28% in mathematics at Grade 4, and 27% in mathematics at Grade 8. Reading was not administered at Grade 8 in 2000, but the percentage of students scoring at the proficient level or above on the 1998 Grade 8 reading assessment was 33%. In mathematics these percentages have improved since 1990, when the percentage of students scoring at the proficient level or above was 13% at Grade 4 and 15% at Grade 8. Thus, the annual gain in percent proficient or above in mathematics averaged 1.5% per year at Grade 4 and 1.2% per year at Grade 8. The gains in

percent proficient or above in reading were much more modest, averaging only three eighths of one percent per year at Grade 4 between 1992 and 2000 and only two thirds of one percent per year at Grade 8 between 1992 and 1998.

The proficient level on NAEP is the one identified by the National Assessment Governing Board as the level that all children should achieve. Although it is easy to agree that this would be desirable, it is also clear that it is extremely ambitious, so much so that it is quite unlikely to be achieved within the foreseeable future, much less by 2014. There is fairly substantial variability from state to state in the percentage of students who score at the proficient level or higher, but in no state has the percentage reached even 50%.

When the NAEP achievement levels were set, the standard setters did not know that the levels might be used to set targets for schools and states that would have rewards and sanctions attached to them. Consequently, it seems problematic to introduce that kind of use after the fact. Moreover, the NAEP achievement levels have been sharply criticized by several national panels of both the National Academy of Education and the National Research Council that have been asked to evaluate NAEP (e.g., Pellegrino, Jones, & Mitchell, 1999; Shepard, Glaser, Linn, & Bohrnstedt, 1993). In addition to finding fault with the process used to set the NAEP performance achievement levels, the National Academy of Education panel concluded that the "achievement levels were set unreasonably high" (Shepard et al., 1993, p. 123).

Although proficient is the label used in NCLB for the level of performance targeted to reach 100% by 2014, proficient is not specifically defined. The label is the same as the name used by the National Assessment Governing Board for the level desired for all students, and thus might be presumed to correspond to the label used in NCLB. As we have seen, however, that level on NAEP appears to be too far out of reach to make a reasonable target that schools and states can realistically aspire to reach in that time frame.

An alternative that would still be quite ambitious, but possibly more within reach with sufficient effort and resources, is the NAEP basic level. The percentages of students in the nation who achieved the basic level or higher on NAEP are displayed in Table 3 by subject and grade for the 1990, 1992, 1996, and 2000 assessments. As can be seen, only in reading at Grade 8 does the percentage approach three quarters of the students. For the other three grade-subject

Table 3

Percentage of Students in the Nation Performing at the Basic Level or Higher on NAEP Reading and Mathematics Assessments by Year

	Reading		Mathematics	
Year	Grade 4	Grade 8	Grade 4	Grade 8
2000	63	NA	69	66
1996	62	74	64	62
1992	60	70	59	58
1990	62	69	50	52

combinations, the percentage is closer to two thirds in the most recent year. Surely, bringing all of the grade-by-subject combinations close to 100% by 2014 would be a major educational accomplishment. Judging by the very small changes in reading from the earlier assessments to the most recent assessment, this may be more than can be reasonably expected in reading. The gains in mathematics are more substantial, but even so, the rate of increase from year to year would have to accelerate to bring the trajectory to the 100% mark in 2014.

If the percentage of students within each state who achieved at the basic level or higher on NAEP were used as a benchmark against which state standards of performance would be compared, it could provide a means of assuring that state standards were less disparate than they now are. At the very least, states with standards that had more students at the proficient level than had reached the basic level on NAEP might be required to provide a rationale to defend their levels.

Index Scores

Attending only to a single cut score, whether it is at the proficient level or at the basic level, gives schools, districts, and states credit for increases in performance only when students make it past the cut score. This narrow focus does not give schools credit for increases in student achievement that occur in the broad range either below or above the cut score. Schools that serve students where the vast majority of students score far below the cut score in a given year might make great improvements in student learning that show up in only a small fraction of the students scoring above the cut score the following year. Substantial increases in the percentage of students who are near the cut score or who are performing better than

their peers the previous year but still considerably below the cut score go unrecognized if only increases in the percent above the cut score are credited.

The NCLB law appears to leave open the possibility that index scores might be used to monitor progress. An index score might, for example, give students who score in the proficient range a score of 1.0, those that score in the high end of the basic range a score of 0.8, those in the mid part of the range 0.6, and those in the low end of the basic range 0.4. Students below basic would receive scores of 0, and those scoring at the advanced level might receive a score of 1.2. The target for such an index score could be an average index score of 1.0 by 2014.

The number of score regions that receive differential values for the index scores, as well as the numerical values to assign to students scoring in those regions, are worthy of empirical analyses to evaluate the properties of the potential index scores. The results of such analyses could help inform deliberations that could be undertaken within a state to choose an index score that would best serve the educational goals of the NCLB law.

Dividing the score scale on a test into regions that are then assigned labels, such as basic, proficient, and advanced, of course ignores differences in performance within each region. Gains in scale scores within a region go undetected and receive no credit. Neither are declines in scores within a region detected. Changes in mean scale scores, on the other hand, would credit improvements in scores anywhere along the scale score and are influenced by all changes in scores, whether positive or negative. Furthermore, there are well-established statistical techniques that might be used to set AYP objectives, such as the use of effect-size statistics that compare differences in means to the standard deviation of scores within a year. For example, the AYP target might be set equal to an annual effect size of .05, that is, an annual increase in the average score equal to .05 standard deviations. Although such an approach may be viewed as out of step with the current emphasis on performance standards, it would have a number of advantages, including improved statistical properties and the crediting of all score changes, not just those that cross the cut score between two performance categories. Yet another advantage of using effect-size statistics is that it would avoid the need to set performance standards, and thereby sidestep the challenge of judging the comparability of performance standards of different states.

One difficult area to balance occurs in the case where new components, such as tests in different content, are added successively to an accountability index. There is tension, on the one hand, in maintaining a system that is intuitively understandable to the policymakers and to educators. However, the appropriate integration and weighting of new tests as they are added to the accountability index is likely to require complex statistical decisions that necessarily reduce the weights of earlier accountability components.

System Validity

The challenge before us is the implementation of legislative intent in a way that will provide the information needed to assess and improve educational quality, information that must be simultaneously relevant to teachers, administrators, policymakers and, of course, parents and students. We have focused on a subset of concerns here that will play out in ways that are appropriate to individual states' and districts' traditions and capacity. Of key importance is to identify the markers and the scientifically based analyses that will provide states and districts with feedback about the utility of their systems. States themselves need to invest in continuing studies (as some of them have) of the impact of their accountability model and the details of its implementation in order to increase the chances that the operations of accountability yield the desired outcome of higher quality education and significantly improved preparation of students.

Conclusion

The NCLB law was motivated by a widely shared desire to improve the education of the nation's youth. Consistent with legislation adopted in many states, the NCLB relies on assessment and accountability requirements as a major mechanism for bringing about desired improvements in student achievement. The accountability requirements go further than the laws in most states in prescribing extensive testing and in setting ambitious objectives for rapid increases in student performance, with the goal that all students achieve at the proficient level or higher by 2014. By requiring that progress be made for subgroups of students defined by race/ethnicity and by economic background, the NCLB also demands more than the current laws in most states.

The NCLB goals are laudable, but the requirements of the law pose substantial challenges for schools, districts, and states. Given the diversity in state content

standards, the rigor of state tests, and the stringency of the cut scores that determine whether students have met the state standard, states will be starting at quite different positions and will vary greatly in how stringent the AYP objectives are unless the law is implemented in a way that makes allowances for the great variability among states in their current testing programs and performance standards. State results on NAEP provide the best source of information that could be used to make such allowances. However, the proficient level on NAEP is set at a level that is to high to hold as a reasonable expectation for all students. The basic level on NAEP is high enough to pose a substantial challenge for schools, districts, and states, but is one that would at least be in the realm of the possible.

Interpretations of the law also recognize the volatility in school-level results from year to year and provide states with latitude to identify ways of reducing that volatility. Possibilities worthy of consideration include the use of index scores, composites across grades, and rolling averages. Potential advantages of working with scale scores and monitoring changes in average scores over time in terms of standard deviation units, thereby avoiding the need for performance standards altogether, also seem worthy of exploration and comparative analyses.

References

Cronbach, L. J., & Furby, L. (1970). How we should measure change: Or should we? *Psychological Bulletin, 74,* 68-80.

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

Elementary and Secondary Education Act of 1965, Pub. L. 89-10.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Kane, T. J., & Staiger, D. O. (2001). *Volatility in school test scores: Implications for test-based accountability systems.* Paper presented at a Brookings Institution Conference.

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

Linn, R. L., & Haug, C. (2002). *Stability of school building accountability scores and gains* (CSE Tech. Rep. No. 561). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

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

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement. Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels.* Stanford, CA: Stanford University, National Academy of Education.

The White House. (2001). *No child left behind.* Washington, DC: The White House.