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

by Robert L. Linn, Eva L. Baker, and Damian W. Betebenner

The No Child Left Behind Act of 2001 substantially increases the testing requirements for states and sets demanding accountability standards for schools, districts, and states with measurable adequate yearly progress (AYP) objectives for all students and subgroups of students defined by socioeconomic background, race–ethnicity, English language proficiency, and disability. However, states' content standards, the rigor of their tests, and the stringency of their performance standards vary greatly. Consequently, the percentage of students who score at the proficient level or higher on the state assessments varies radically from state to state. Some states have farther to go than others to meet the mandated target of 100% proficient within 12 years. These differences are illustrated and the implications for achieving AYP targets are discussed. Also addressed are possible uses of results from the biennial state-level administrations of the National Assessment of Educational Progress as a means of leveling the playing field. Factors contributing to the volatility of gains in achievement from year to year for individual schools are discussed.

By making accountability the centerpiece of the education agenda, President Bush (The White House, 2001) strongly reinforced what was already a central theme of state policies 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 require changes in the practices of many states. The law requires, for example, that states develop both content standards in reading and mathematics and tests that are linked to those standards for 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 test all students in Grades 3 through 8.

Educational Researcher, Vol. 31, No. 6, pp. 3–16

The focus of this article is on key provisions of the law related to performance standards, adequate yearly progress (AYP) targets, and the challenges such targets present both methodologically and practically for states and for schools. Before turning to those issues, however, 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. States vary in the way they expect such assessments to map to standards. In addition to difficulty levels (associated with both the actual items and tasks used on an assessment and the stringency of performance standards), testing programs vary, at least nominally, in 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 (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 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, AYP, and an external arbiter for performance (e.g., the National Assessment of Educational Progress) needs to be considered in 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 these changes are those concerned with the identification of AYP 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.

NCLB has immediate implications for states that must put in place new testing and accountability systems. Over the next several years, however, the requirements of NCLB have implications for all educators and educational researchers who focus on

K–12 education. The implications for teachers and school administrators derive from the requirements of the law that schools demonstrate steady gains in student achievement and close the gap in achievement between various subgroups of students. Schools that fail to meet improvement targets must adopt alternate instructional approaches or programs that have been shown to be effective through *scientifically based research,* a phrase that appears 111 times in the NCLB law (Feuer, Towne, & Shavelson, in press). NCLB clearly presents a major challenge to the field of educational research.

In the following sections we focus on critical requirements of the law. We begin by summarizing the requirements to show AYP. Because performance standards are central to the notion of AYP but not well defined by NCLB, we turn to a relatively extended discussion of the between-state variation in performance standards and the implications of that variation for AYP targets. Subsequent sections consider issues related to the establishment of AYP objectives, implications of the requirements for individual schools' results, potential uses of results from the required biennial NAEP, and some alternatives to tracking the percentage of students who score at the "proficient" level or above. The challenges posed by the NCLB law are many; unless considerable flexibility is allowed in the interpretation of some aspects of the accountability components of the law, it seems likely that many more schools will be placed in the improvement category than can be provided with effective assistance. Such an outcome could seriously undermine the law's laudable goals of substantial improvement in instruction and learning for all students and closing the achievement gap.

Adequate Yearly Progress

NCLB specifies that states must develop AYP objectives consistent with the following requirements 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. 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 requirement—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 established performance standards for their tests, the standards were set unaware that they would be used to determine AYP objectives or that substantial sanctions would be associated with failure to meet AYP targets. Instead, panels of teachers—either alone or with other interested citizens—have generally set performance standards. The panels review tests and identify cut scores thought to correspond to the level of performance expected from a proficient student who is motivated to do well and has had an adequate opportunity to learn the material.

The result of this judgmental standard setting process frequently has been to set the proficient level so high that it may be unrealistic to expect all students to reach that level by 2014. Certainly, no state, or country for that matter, is close to meeting the high standard set for proficient performance on NAEP or similar standards on many state assessments (Linn, 2000). Indeed, only a very few schools with student bodies selected on the basis of past achievement or from privileged backgrounds now have all their students scoring at the elevated proficient levels of the more rigorous state tests.

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 that states have set, which determine the cut scores used to define proficient performance, also vary widely from one state to another. For example, the percentage of students reported on the respective state department of education websites to have scored at the proficient level or higher in 2001 on the state Grade 8 mathematics assessments was 39% in Mississippi and only 7% in Louisiana. The percentage of students who "passed" the Grade 8 mathematics assessment in Texas in 2001 was 92%. Although there may be real differences in mathematics achievement in these three states, those differences certainly are not as great as the differences in these percentages. Clearly, *proficient* or *passing* have quite different meanings in these three states.

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] The five states for Figures 1 and 2 were selected to illustrate a range of types of tests, uses of test results, and performance standards. As can be seen, although 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 the standard on the state Grade 8 reading tests ranged from 27% to 91%. The corresponding range for the Grade 8 mathematics tests was from 31% to 92%. A straight-line projection of gains needed between 2001 and 2012 would require an annual

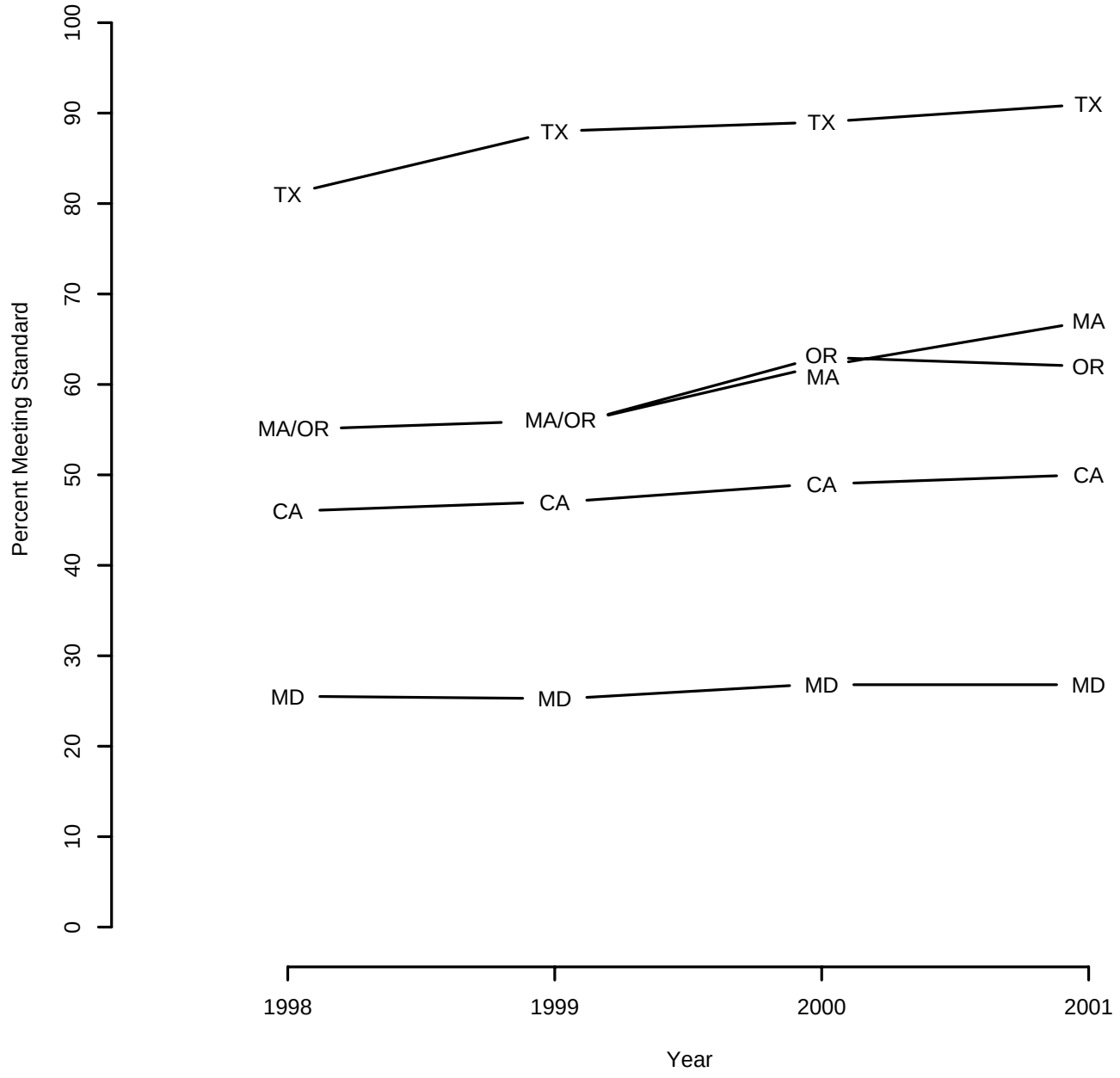

FIGURE 1. *Trends in percentage meeting standard in five states: grade 8 reading.*

increase of slightly less than 1% per year for the state with the highest percentages meeting standards in 2001 and more than 5% per year for the states 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 in student achievement are available for those two states for the 8 years starting in 1994 and ending in 2001. Maryland and Texas also participated in the state-level administration of the Grade 8 NAEP mathematics assessments in 1990, 1992, 1996, and 2000.

The trend in percentage of students passing the Grade 8 Texas Assessment of Academic Skills (TAAS) is plotted from 1994 to 2001 in Figure 3. Also shown in Figure 3 are the trends in the percentage of students who scored at the basic level or higher and the percentage of students who scored at the proficient level or higher on NAEP. As can be seen, there was a substantial increase in the percentage of students who passed the TAAS Grade 8 mathematics assessments over the 8-year period from 1994 through 2001. There was also an upward trend in the percentage scoring at the basic level or above and at the proficient level or above on NAEP. The slopes of the trend lines on NAEP are not as steep, however, as the slope of the TAAS trend line.

There are, of course, differences between TAAS and NAEP that may contribute to the difference in levels and slopes of the trend lines in Figure 3. Although Texas is in the process of introducing new, more demanding tests, the test in place during the years for which results are graphed in Figure 3 primarily measured basic skills; NAEP is a more challenging test that measures more complex reasoning and problem-solving skills. TAAS is

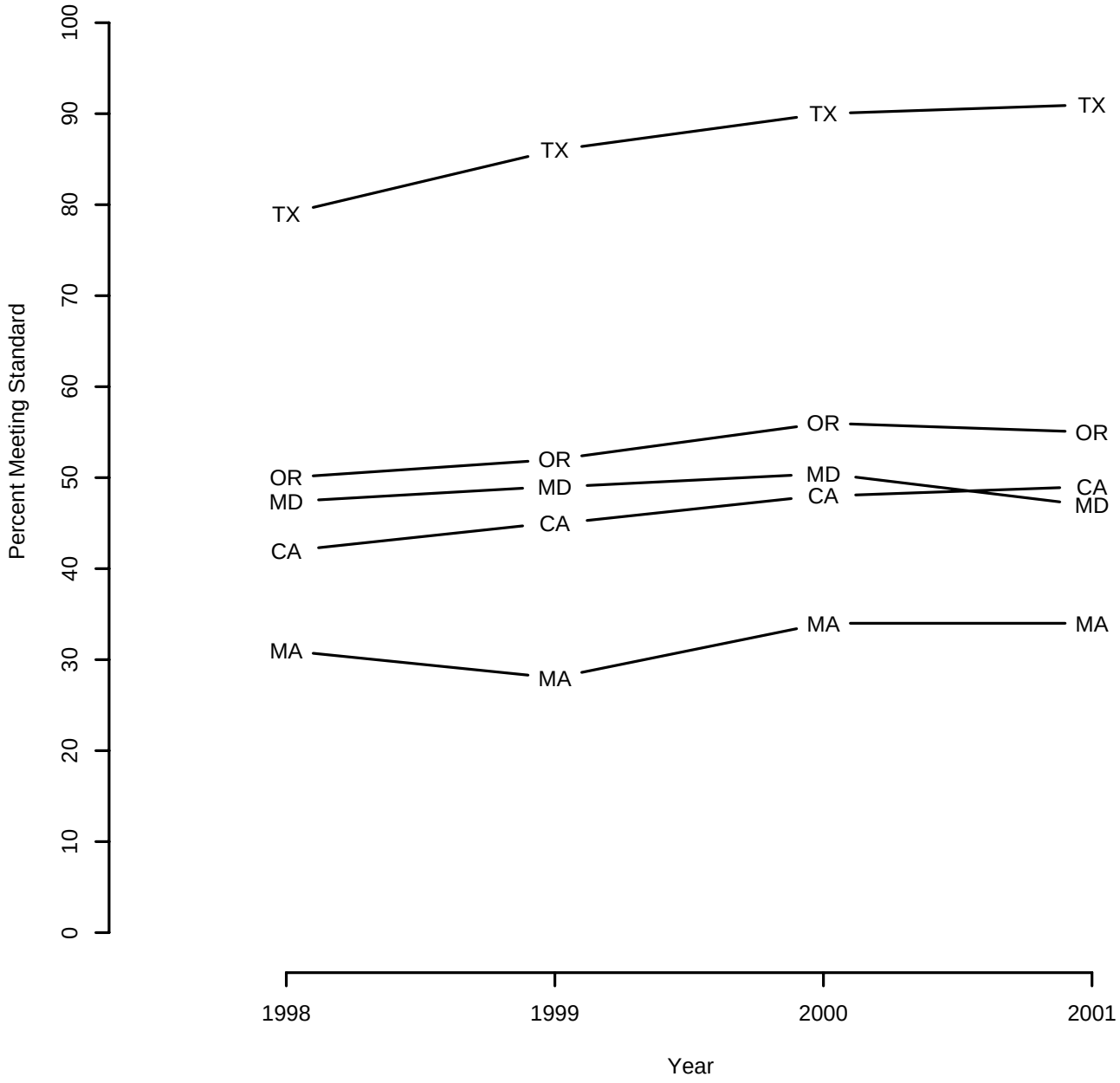

FIGURE 2. *Trends in percentage meeting standard in five states: grade 8 mathematics.*

also a relatively high-stakes test for students, whereas there are no stakes for students associated with their performance on NAEP. Consequently, differences in motivation may play a role in the differences shown in Figure 3. It should be noted, however, that when NAEP has been administered under conditions expected to increase student motivation (e.g., embedding sections of NAEP in a state test, or providing students with incentives to perform well on NAEP), the differences in performance, although statistically significant, have been quite small (Kiplinger & Linn, 1996; O'Neil, Sugrue, & Baker, 1996). Small effects were obtained even when students were paid $1.00 for each correct answer (O'Neil et al., 1996).

Despite differences in the stakes attached to the results of state tests and measures such as NAEP in content coverage, it is relevant to ask the degree to which gains on a state test generalize to gains on other measures of achievement. When there are gains on a state test, are there also gains on another measure of achievement (e.g., on NAEP) in the same content area? This is because of concerns that the narrow focus on teaching to a state test may produce inflated gains in scores and because the fundamental concern is with improved achievement, not just higher test scores (Amrein & Berliner, 2002; Koretz & Baron, 1998; Stecher & Hamilton, 2002).

Figure 4 displays trends in the percentage of students meeting the standard on the Maryland School Performance Assessment Program (MSPAP) in mathematics at Grade 8 from 1994 through 2001. The trends in the percentages of student scoring at the basic level or higher or at the proficient level or higher for the four state-level NAEP mathematics assessments at Grade 8 from 1990 through 2000 are also displayed. The fact that the MSPAP

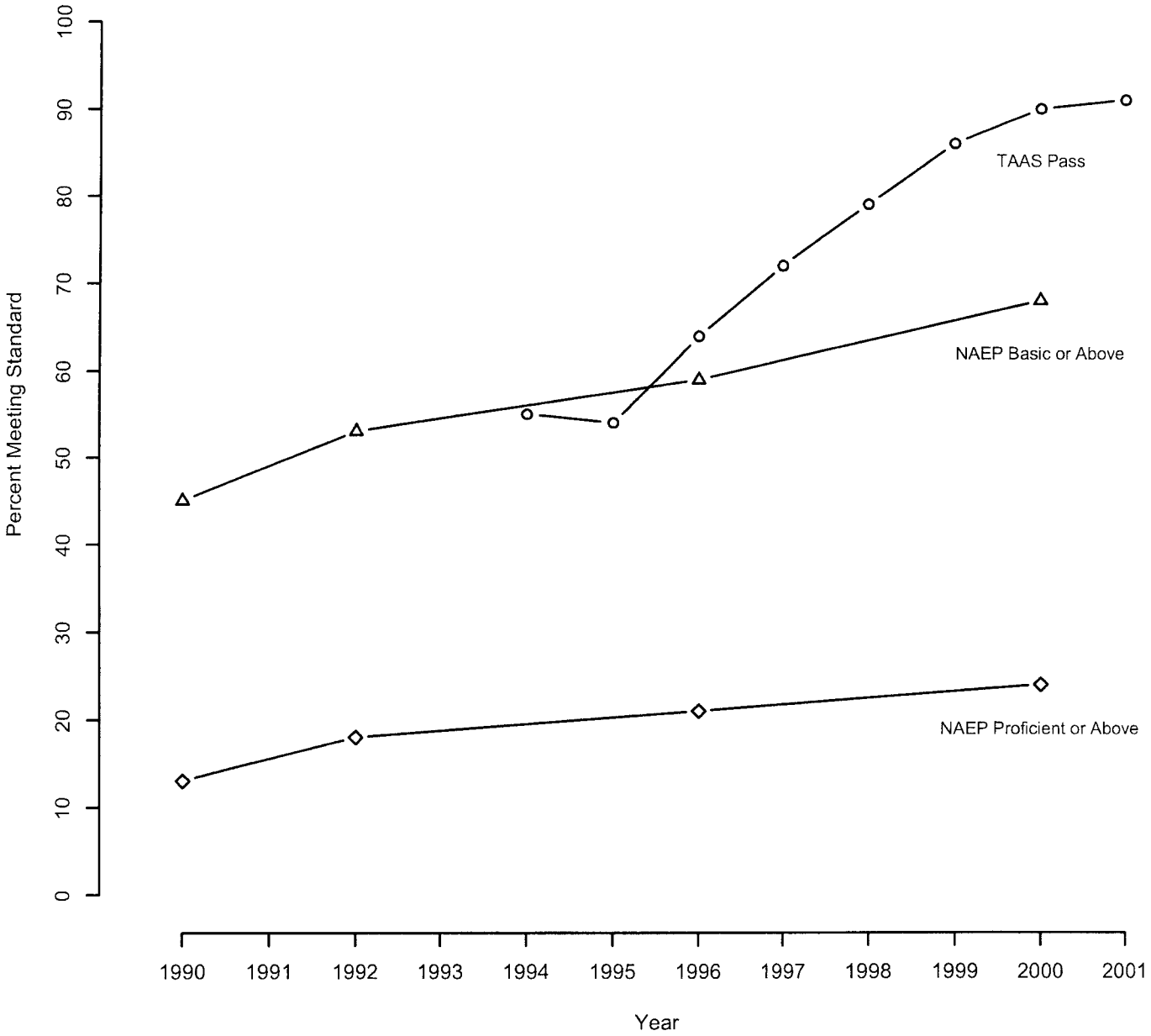

FIGURE 3. *Texas trends in grade 8 mathematics performance on TAAS (percentage pass) and on NAEP (percentage basic or above and percentage proficient or above).*

line falls between the two NAEP lines suggests that the standard to be met on MSPAP is more stringent than the basic level on NAEP but less stringent than the NAEP proficient level. From a comparison of the slopes of the NAEP trend lines for Texas and Maryland in Figures 3 and 4 it can be seen that both the levels and slopes are quite similar for the two states. On the other hand, the trend line for MSPAP has nearly the same slope as the Maryland NAEP trend lines, whereas, as previously noted, the TAAS trend line is steeper than the Texas NAEP trend lines. The similar slopes of MSPAP and NAEP in Maryland may reflect the fact that the content of these two assessments is similar. Both MSPAP and NAEP are reasonably challenging assessments. Furthermore, stakes on the MSPAP are mainly at the school level rather than at the level of individual students. It also may be that it is more difficult to achieve substantial gains on more ambitious tests measuring complex reasoning and problem-solving skills than it is on tests that primarily measure basic skills.

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 NCLB set targets based on schools increasing 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 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.

A steady increase of at least a point per year would still leave states far short of the goal of all students at the proficient level by

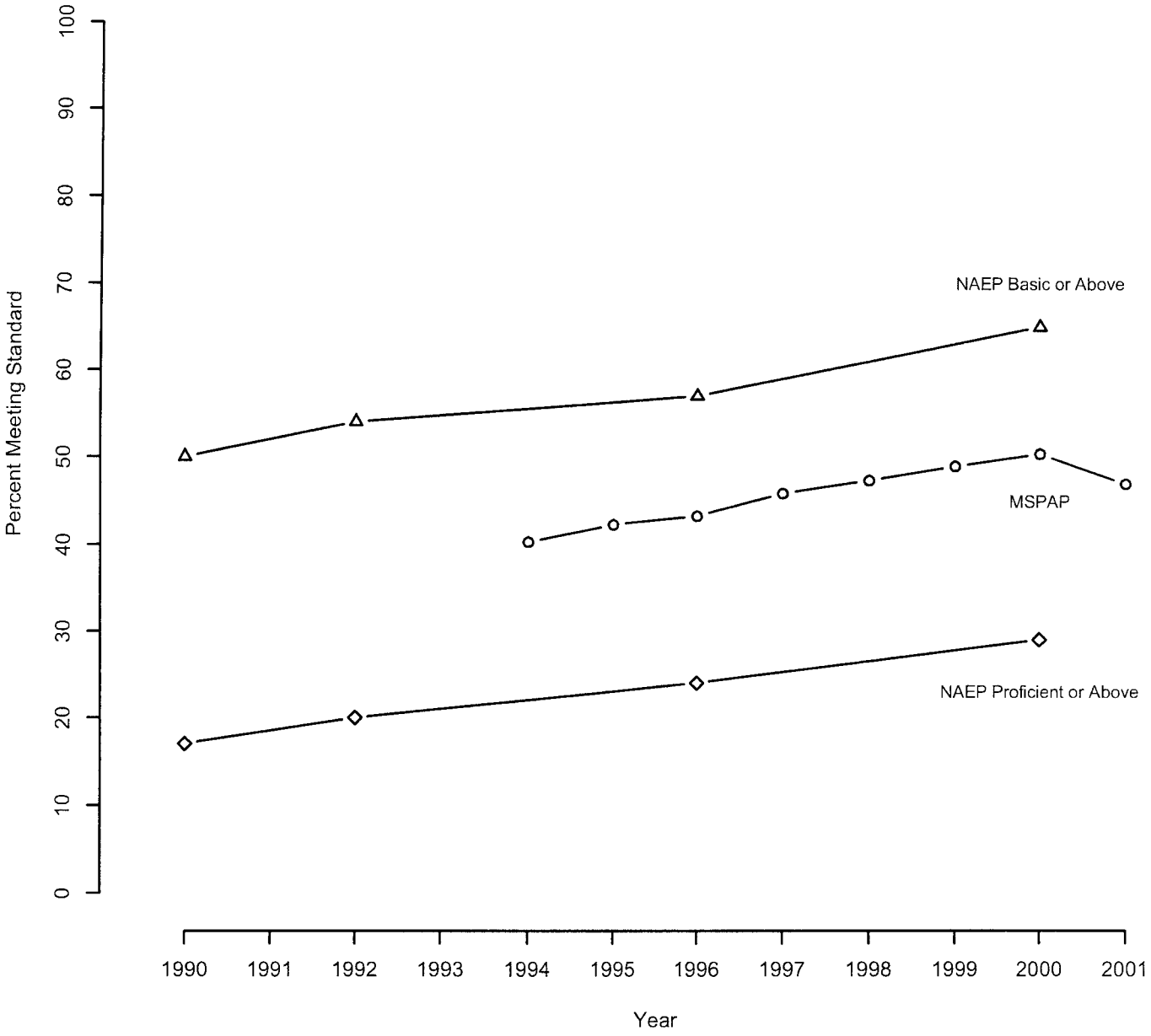

FIGURE 4. *Maryland trends in grade 8 mathematics performance on MSPAP (percentage meeting standard) and on NAEP (percentage basic or above and percentage proficient or above).*

2014, and these targets are unlikely to be met by a large number of 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 change in the percentage of students who score at the proficient level or higher on NAEP has differed by state, but has been relatively modest during the last decade for most states. Figure 5 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 dashed 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 6, states showed larger gains on the Grade 4 mathematics assessments 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 of 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 assessments in both 1990 and 2000 had increases in the percentage of students scoring at the proficient level or higher (see Figure 7). Judging from these

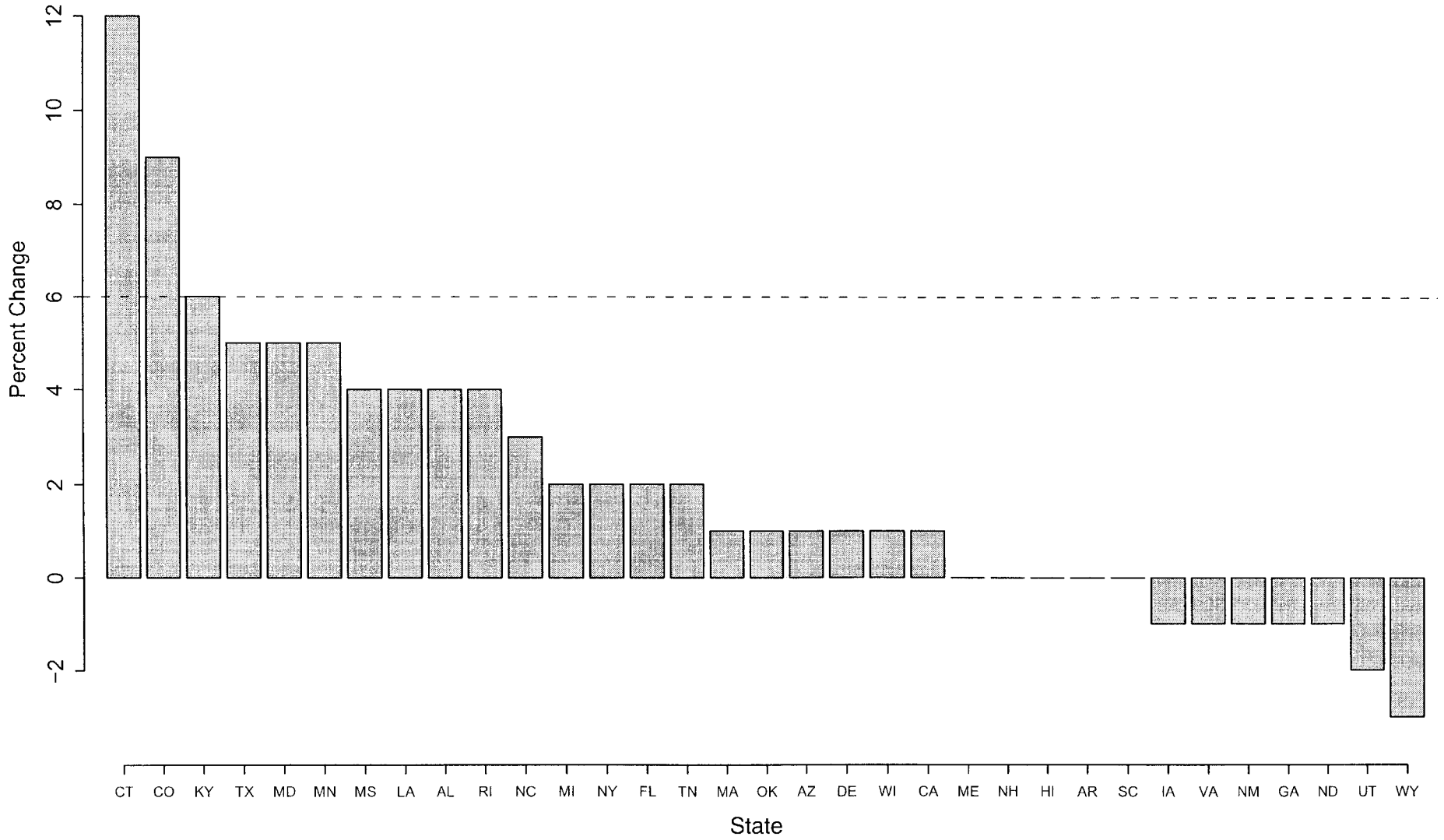

FIGURE 5. *State changes in percentage proficient or above from 1992 to 1998 on NAEP grade 4 reading assessment.*

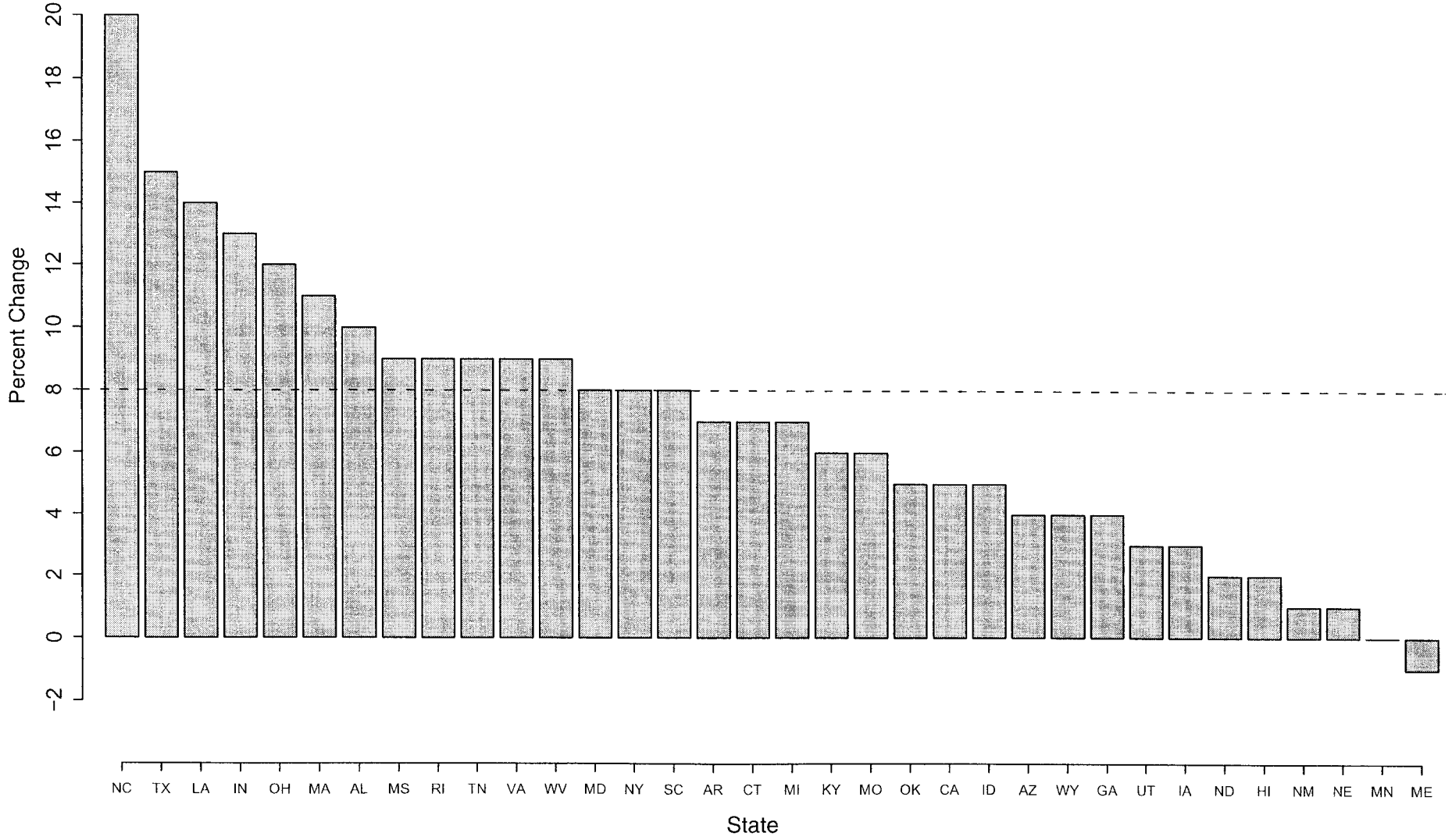

FIGURE 6. *State changes in percentage proficient or above from 1992 to 2000 on NAEP grade 4 mathematics assessment.*

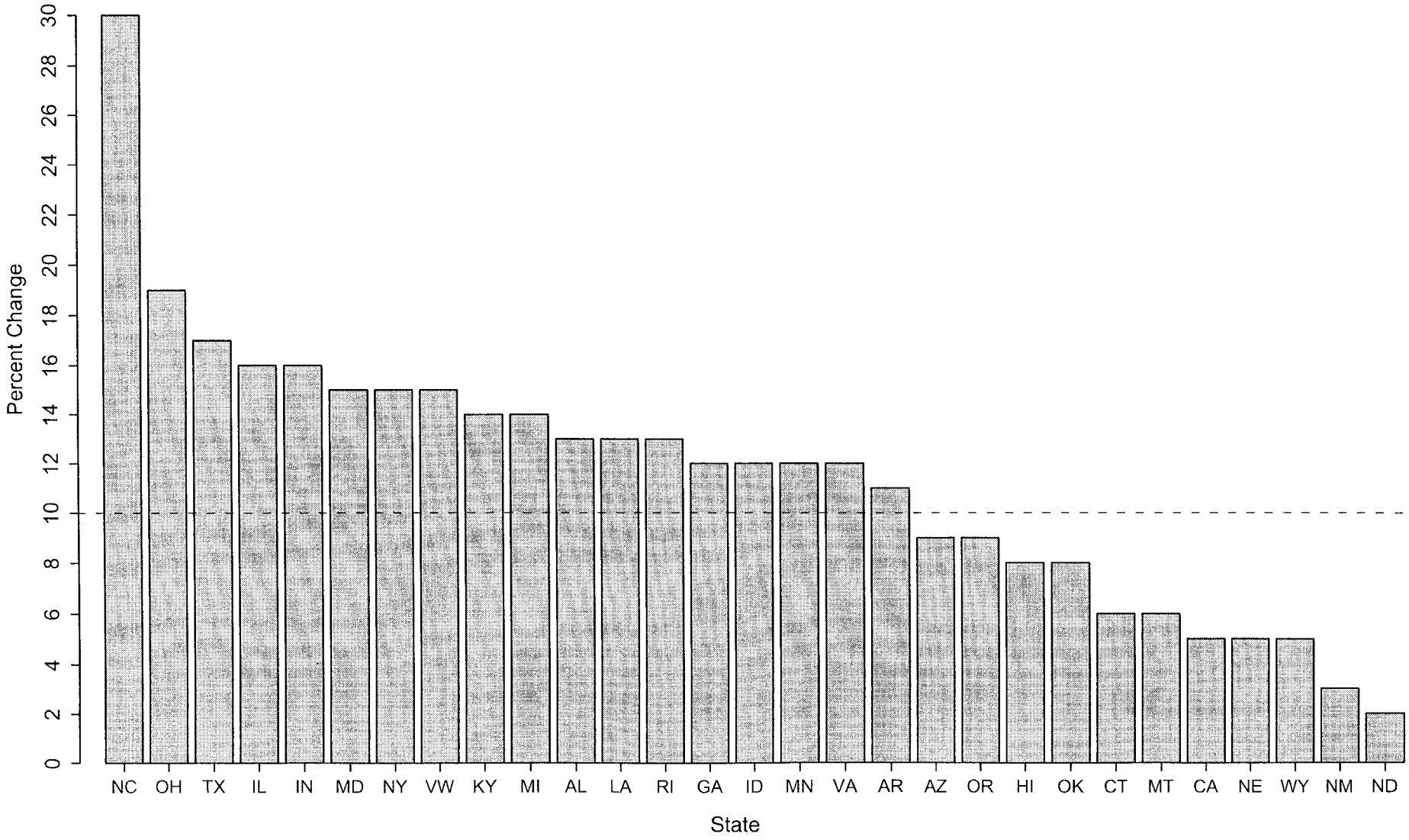

FIGURE 7. *State changes in percentage proficient or above from 1990 to 2000 on NAEP grade 8 mathematics assessment.*

NAEP results, it seems clear all schools showing increases of one point or more in both reading and mathematics is a target that is extraordinarily ambitious. 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 at that level, it is clear many more than 12 years would be required to reach the NCLB goal.

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 many state laws passed in the last few years. 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 can 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, having 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 are unachievable.

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. 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 of this unreliability. School summary scores for each year are subject to not only measurement error but to sampling error as well. 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.

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 students' socioeconomic backgrounds. 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 a target of a 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.

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 students but for every subgroup of students in the school. Furthermore, as was previously indicated, steady progress of a one-point increase per year is not sufficient to bring the

Table 1

Number and Percentage of Schools that Met or Exceeded Target of an Increase in the Percentage of Student 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

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

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 3 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 school 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, as illustrated, the volatility due to sampling error and nonpersistent factors is so great that schools that are 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. Requirement 7—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 AYP 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., five out of eight or seven out of ten 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 requirements 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 states with reasonably ambitious tests and performance standards would have unobtainable AYP objectives. Hence, it is highly desirable that interpretations and guidance from the U.S. Department of Education contribute both to leveling the playing field across states and to making 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 of improving the achievement of all children and closing the gap in achievement among racial and 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 the biennial participation in state NAEP in reading and mathematics at Grades 4 and 8. The ways in which NAEP might be used independently to 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. In principle, there are many ways in which NAEP might be used. At one extreme, for example, state NAEP results could be used to define the percentage of students at Grades 4 and 8 who achieve at various levels. 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, by interpolation at Grades 5, 6, and 7, and by extrapolation at Grade 3.

Table 2

Number and Percentage of Schools Meeting or Exceeding Target of an Increase in the Percentage of Student 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 Tests)

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

Making NAEP the controlling factor would be fair to states; 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 do so for purposes of reporting scores of individual students (Feuer, Holland, Green, Bertenthall, & 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 due to 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 percentages of students scoring at the proficient level or above were 13% at Grade 4 and 15% at Grade 8. Thus, the annual gain in the percentage proficient or above in mathematics averaged 1.5% per year at Grade 4 and 1.2% per year at Grade 8. The gains in the percentage proficient or above in reading were much more modest, averaging only three eighths of 1% per year at Grade 4 between 1992 and 2000 and only two thirds of 1% 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 school and state targets 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 NCBL 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 NCBL. As we have seen, however, that level on NAEP appears 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 still would be quite ambitious but possibly more attainable 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 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 from the very small changes in reading from the earlier assessments to the most recent

Table 3

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

Year	Reading		Mathematics	
	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

assessment, it 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 percentages 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 could be compared, it would assure that state standards were less disparate than they now are. At the very least, states having standards that had more students at the proficient level than at 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 or 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 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 are still below the cut score) go unrecognized if only increases in the percentage above the cut score are credited.

Flexible interpretation of the NCLB law could include the possibility that index scores 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 and the numerical values assigned 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 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 score scale 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 use of 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 are 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 of yielding the desired outcome of higher quality education and significantly improved preparation of students.

Conclusion

NCLB 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. By requiring that progress be made for subgroups of students defined by race, ethnicity, and economic background, the NCLB again 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 state cut scores, states will be starting at quite different positions. States' AYP objectives will vary in stringency 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 too high to be held 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 would at least be in the realm of the possible.

Interpretations of the law also should 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.

NOTES

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. The findings and opinions expressed in this report do not reflect the positions and policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

[1] California uses the SAT9. The 50th national percentile rank was used as the cut point for the graphs. Maryland uses the MSPAP. The percentage of students scoring satisfactory or better is plotted in the graphs. Massachusetts uses the MCAS. The percentage of students scoring proficient or better is plotted in the graphs. Oregon uses the Oregon Statewide Assessment. The percentage of students meeting or exceeding performance standards is plotted in the graphs. Texas uses the TAAS. The percentage of students meeting the minimum expectations is plotted in the graphs.

[2] It is a surprise to some that sampling error is relevant because 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).

REFERENCES

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Archives*, *10*(18), 1–70.

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., Bertenthall, M. W., & Hemphill, C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Feuer, M. J., Towne, L., & Shavelson, R. J. (in press). Scientific culture and educational research. *Educational Researcher*.

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.

Kiplinger, V. L., & Linn, R. L. (1996). Raising the stakes of test administration: The impact on student performance on NAEP. *Educational Assessment*, *3*, 111–133.

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

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *23*(9), 4–16.

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

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.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1996). Effects of motivational interventions on NAEP mathematics performance. *Educational Assessment*, *3*, 135–157.

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: A report of the National Academy of Education Panel of the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels*. Stanford, CA: Stanford University, National Academy of Education.

Stecher, B. M., & Hamilton, L. S. (2002). Putting theory to the test: Systems of "educational accountability" should be held accountable. *Rand Review*, *26*(1), 17–23.

The White House. (2001). *No Child Left Behind.* Washington, DC: The White House.

AUTHORS

ROBERT L. LINN is a distinguished professor at the University of Colorado, Boulder School of Education, Campus Box 249, Boulder, CO 80309-0249; Robert.Linn@colorado.edu. He is also Co-Director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His research interests include educational measurement, and the design and impact of educational accountability systems.

EVA L. BAKER is a professor at the University of California, Los Angeles and Co-Director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), GSE & IS Building, Los Angeles, CA 90095-1522; baker@gseis.ucla.edu. Her research interests include assessment policy and technology-based systems for assessment and learning.

DAMIAN W. BETEBENNER is a research associate at the University of Colorado at Boulder, School of Education, 249 UCB, Boulder, CO 80309-0249; damian.betebenner@colorado.edu. His research interests include the use of large-scale data sets to investigate the efficacy of educational policy initiatives involving accountability and school choice.

Manuscript received March 13, 2002
Revisions received May 9, 2002
Accepted May 11, 2002