Accountability:
Responsibility and Reasonable Expectations

CSE Report 601

Robert L. Linn
CRESST/University Colorado at Boulder

July 2003

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 1.1: Comparative Analyses of Current Assessment and Accountability Systems. Strand 2: Outcomes of Different Accountability Designs
Project Directors: Robert L. Linn, Lorrie Shepard, and Haggai Kupermintz, CRESST/University of Colorado at Boulder, and Meredith Phillips and Eva L. Baker, CRESST/UCLA

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Institute of Education Sciences, U.S. Department of Education.

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

ACCOUNTABILITY: RESPONSIBILITY AND REASONABLE EXPECTATIONS[1]

Robert L. Linn
CRESST / University of Colorado at Boulder

Abstract

Some of the central features of current educational accountability systems are discussed using the requirements under the No Child Left Behind (NCLB) Act of 2001 as the primary example. It is argued that broadly shared responsibility is needed for accountability systems to contribute to improved education. It is also suggested that systems need to be designed in ways that are consistent with research and past experience. This requires the setting of ambitious performance standards and improvement targets, but ones that can reasonably be achieved given sufficient effort and supporting resources. These design features are contrasted with the NCLB requirements. Illustrations are provided of some of the state responses to the NCLB demands that attempt to avoid the over-identification of schools for improvement and sanctions.

Demands for greater accountability for student achievement are an all too familiar part of political statements and legislative actions concerning education. Accountability, as mandated in federal and state legislation, is intended to improve the quality of education for all students. No reasonable person is against accountability that enhances the quality of education. But reasonable people can and do differ regarding what an accountability system needs to entail if it is to help us achieve the shared goal of improved education. It is my contention that, among other things, accountability must entail broadly shared responsibility if it is going to have the positive effects that it is expected to have without having unintended negative effects.

Shared Responsibility

When a group of educators in Washington State were asked what words or concepts they thought should be associated with accountability, the most frequent word was "responsibility" and the second most frequent was "shared" (T. Bergeson,

[1]Presidential Address, American Educational Research Association, Chicago, IL, April 23, 2003. I thank Eva Baker, Joan Herman, and Lorrie Shepard for helpful comments on a draft of this paper.

personal communication, 2001). Shared responsibility was viewed more broadly by the Washington educators than it seems to be in most laws that have mandated accountability systems in recent years. Shared responsibility was broadly conceived to include students, teachers, school administrators, parents, and policymakers.

Most state and federal laws that have been enacted in the past few years, as well as policy discussions of accountability, have focused more narrowly on educators and students. Both educators and students clearly need to share part of the responsibility, but so must administrators, parents, and policymakers. Students and teachers have a responsibility to put forth a reasonable level of effort, while administrators and policymakers have a responsibility to provide the means—both instructional resources and professional development—for teachers and students to meet the expectations of the accountability system. Parents also need to share responsibility.

Porter and Chester's (2001) discussion of "symmetric accountability" systems provides a rationale for the belief that shared responsibility of students, teachers, administrators and policymakers is critical for accountability to contribute to improvements in education. Their idea of symmetry is that incentives need to be distributed across all parties to encourage the shared responsibility that is needed for real improvement. Although broadly shared responsibility is the most likely way for expectations to be achieved, the reality is that most accountability systems now in place focus so heavily on educators and/or students that others are largely ignored. Greater emphasis needs to be given to other responsible parties.

Researchers also need to share responsibility. We have the responsibility to provide solid information about the strengths and weaknesses of alternative approaches and interventions, one of which is the accountability system itself. I doubt that anyone would say that we already know all we need to know to design a highly effective accountability system that is sure to contribute to the broad goal of improving education without having major unintended negative side effects. However, we do have a good deal of information based on research and past experience with a variety of accountability systems that have been used by states and the federal government. Accountability systems need to be designed in ways that are consistent with that evidence and past experience regarding the factors that enhance positive effects and minimize negative effects.

Key Elements of Accountability Systems

What Counts

Two questions that need to be considered for any accountability system are (a) What counts? and (b) Who is held accountable? As was implied above, the "who is accountable" question is often answered too narrowly. Past research evidence and experience suggest that the "what counts" question is also generally answered too narrowly. The measures that enter into the accountability system should be broadly conceived and provide information on a wide range of outcome, contextual, and process variables. The range of outcome measures needs to be broad to avoid overemphasis of some goals at the expense of others. Data on contextual and process variables are needed to interpret results on the outcome measures and to suggest desirable directions for change. If it were known, for example, that few teachers had the needed preparation in mathematics, that information would be relevant both for interpreting lackluster student mathematics achievement and for suggesting actions that are likely to lead to improvement.

Unfortunately, the laws and regulations that have mandated accountability systems have too often defined the measures that count in far too narrow a way. Consider, for example, the No Child Left Behind (NCLB) Act of 2001 (Public Law 107-110), reauthorizing the Elementary and Secondary Education Act (ESEA) of 1965. NCLB has much that is worthy of praise. It stays the course on standards-based reform and encourages states to adopt ambitious subject-matter standards. It is also praiseworthy for the emphasis on all children and the particular attention it gives to promoting the learning of groups of students that have lagged behind in the past. NCLB also contains a number of questionable features, however, one of which is the narrow definition of what counts. The student measures that count in NCLB are dominated by tests of reading/language arts and mathematics. It is true that science will be added later at selected grades. Requirements including high school graduation rates, the proportion of students tested, and other academic indicators are required at all grade levels, but results on each state's reading/language arts and mathematics tests dominate all other indicators. Indeed, if a state elects to use other indicators, NCLB explicitly prohibits the state from using "those indicators to reduce the number of, or change the schools that would otherwise be subject to school improvement, corrective action, or restructuring" (P.L. 107-110, sec. 1111, (b)(2)(A)(i)).

It is no surprise that attaching high stakes to test results in an accountability system leads to a narrowing of the instructional focus of teachers and principals. There is considerable evidence that teachers place greater emphasis on material that is covered on a high-stakes test than they do on other material (e.g., Mehrens & Kaminski, 1989; Shepard, 1990; Stecher & Hamilton, 2002; Taylor, Shepard, Kinner, & Rosenthal, 2001). Within a content area, this concentration of effort has positive as well as negative aspects. The concentration is desirable if it leads to greater emphasis on the knowledge and skills stressed in challenging content standards that the test is intended to measure. The concentration can be negative, however, when it focuses too closely on the specific item formats and parts of the content domain that are emphasized to a greater degree on the test without attention to broader subject-matter standards (Shepard, 1990, 2000). Narrowing the focus within a content area to material tested can result in an impoverished definition of reading, writing, or mathematics.

Furthermore, the concentration on tested content areas often comes at the expense of content domains that are not tested, such as science, history, geography, and the arts (e.g., Stecher & Hamilton, 2002; Taylor et al., 2001). It is in recognition of the potential to narrow the curriculum to those content areas that are tested that some accountability systems have sought to include a wide range of content areas. The Kentucky accountability system, for example, includes tests at selected grades in seven content areas (reading, writing, mathematics, science, social studies, arts and humanities, and practical living/vocational studies; http//www.kde.state.ky.us/). Concerns about overemphasis of tested subjects at the expense of other subjects also explains why several states have found that teachers specializing in subjects such as science, history, or the arts are eager to have the state include tests of those subjects in their accountability system. Including tests in a subject area in the accountability system is certainly not the only way to assure adequate attention being given to the subject, but including tests of the subject is seen as one way of increasing attention given to the subject.

Expectations

Objectives mandated by an accountability system should be ambitious, but also should be realistically obtainable with sufficient effort. The question is how do you know whether ambitious goals are realistic? Are they not being achieved because of insufficient effort? Are they not being achieved because insufficient resources are being devoted to the problem? I believe that past experience is the first place to look.

It is not that current levels of student performance, or gains in student performance that typically have been achieved in the past, are fine and should be adopted as the standard to be expected in the future. Rather, current levels of performance and past gains provide a context for judging future gains and longer range targets of performance. At the very least, there should be what I call an existence proof. That is, we should not set a goal for all schools that is so high that no school has yet achieved it. For example, if no school has 100% of its students scoring at the proficient level or higher, we should not expect all schools to reach that level in the next 12 years. Indeed, I would argue that if 90% of the schools currently fall short of the 100% proficient goal, then that is an unrealistic expectation for all schools to achieve within a dozen years.

Establishing student achievement expectations involves at least four critical components. First, there are the questions about how the content domains should be defined. Second, there are questions about how the identified content domains should be assessed. Third, given the current emphasis on reporting in terms of performance standards, referred to by NCLB as "academic achievement standards," there are questions about the establishment of the performance standards. Finally, there are questions about establishing long-range goals and intermediate performance objectives.

Content and Performance Standards

The No Child Left Behind Act of 2001 addresses all four of the above critical aspects of setting expectations, albeit, as would be expected of legislation, only in quite general terms. NCLB provides the following guidance in delegating to states the responsibility of setting content and student performance standards. "Each State shall demonstrate that the State has adopted challenging academic content standards and challenging student academic achievement standards that will be used by the State" (P.L. 107-110, Section 1111(b)(1)(A)).

Almost every state has adopted content standards for the areas of reading/language arts and mathematics. They vary greatly in their specificity and arguably in how challenging they are, but each state will be able to claim that their content standards are in compliance with NCLB. Regardless of the strength or weakness of the claim, it is unlikely that a state's content standards will be found to be out of compliance because the federal government is reluctant to get into the business of specifying content coverage.

Most states have also set student performance standards. A number of states have used the National Assessment of Educational Progress (NAEP) as their model in setting performance standards for their state assessments. NAEP performance standards divide the range of scores into four achievement levels, called below basic, basic, proficient, and advanced. Although a number of states define performance standards that also distinguish four levels of performance, the labels attached to those levels vary. Labels such as "exceeds standard," "superior," and "distinguished, as well as "advanced," are used by various states for their highest level of performance. As is discussed in greater detail below, the state performance standards also differ in stringency from state to state and in comparison to NAEP.

The NAEP achievement levels are quite ambitious performance standards. The ambitious nature of the NAEP standards has been discussed in some detail in several reports (e.g., Pellegrino, Jones, & Mitchell, 1998; Shepard, Glaser, Linn, & Bohrnstedt, 1993). The percentage of students scoring at the advanced level and the percentage of students scoring at the proficient level or higher on NAEP from the most recent assessments in reading and mathematics are shown in Table 1. A simple inspection of the percentages of students performing at either the advanced or proficient levels provides an indication that the standards are set at high levels.

For the results displayed in Table 1, the NAEP proficient standards in reading are set at close to the 70th percentile at Grades 4 and 8 and at the 60th percentile at grade 12. The standards are even more stringent in mathematics where the proficient standard is set at nearly the 75th percentile at Grades 4 and 8 and a little higher than the 80th percentile at Grade 12. Students who perform at the advanced level on NAEP are quite rare in both subjects. The demanding nature of the NAEP

Table 1

National Percentages of Students Scoring at the Proficient Level or Above and Percentages Scoring at the Advanced Level on NAEP in 2000

	Percent Proficient or Above		Percent Advanced	
Grade	Reading	Mathematics	Reading	Mathematics
4	32	26	8	3
8	33[a]	27	3	5
12	40[a]	17	6	2

Source: http://nces.ed.gov/nationsreportcard/

[a]Grade 8 and 12 Reading results are for the 1998 assessment.

performance standards is also apparent from linkages of NAEP to the Third International Mathematics and Science Study (TIMSS) Grade 8 mathematics results, which reveal that no country is anywhere close to having all of its students scoring at the proficient level or higher (Linn, 2000).

NAEP performance standards are clearly ambitious, maybe too ambitious. Certainly the target of 100% proficient or above according to the NAEP standards appears more like wishful thinking than a realistic possibility. Nonetheless, the NAEP achievement levels provide a benchmark against which the relative stringency of standards set by various states on their own assessments can be compared. McLaughlin and Bandeira de Mello (2002) have linked state assessment data to NAEP data for a number of states that have participated in state NAEP. Through the linkages they are able to compare the performance standards set on state assessments to the NAEP performance standards (proficiency levels). For example, McLaughlin and Bandeira de Mello provide a comparison of the performance standards for the Grade 4 mathematics assessments used in 2000 for 19 states. They provide a chart in which state standards are plotted on a common scale together with the NAEP achievement levels. The chart includes multiple performance standards for most states. The highest level identified by the 19 states is never as stringent as the NAEP advanced level, though Maine's highest level, called "exceeds standard," comes close.

The highest levels of eight other states fall somewhere between the NAEP advanced and proficient standards. The "advanced" performance standard in three of those states, Louisiana, Wyoming, and Massachusetts, is closer to the NAEP advanced cut than it is to the NAEP proficient cut. The Kansas, Missouri, and South Carolina advanced standards are about equidistant from the NAEP advanced and proficient cuts, whereas the Georgia and New York "exceeds standard" cuts are close to the NAEP proficient cut.

At the other end of the scale, McLaughlin and Bandeira de Mello's (2002) results show that the Georgia "meets standard" and the North Carolina "consistent mastery" levels are less stringent than the NAEP basic level, whereas the New York "meets standard" and the Connecticut "goal" levels are slightly more demanding than the NAEP basic level. Overall, McLaughlin and Bandiera de Mello's comparisons of the state performance standards to NAEP standards clearly indicate that there is considerable variability across states in the stringency of their standards.

Use of Performance Standards in Setting Improvement Objectives

NCLB includes the requirement that states participate in biennial administrations of state NAEP in reading and mathematics at Grades 4 and 8 starting in 2003. The law does not indicate what use, if any, will be made of the NAEP results. There is a vague notion, however, that NAEP will provide some kind of benchmark against which state results can be compared. From the discussion above, it is clear that NAEP could be a way of checking on the stringency of state performance standards. It also could be useful as a way of judging the trustworthiness of trends in achievement that states report on their assessments over the next several years (see, for example, Linn, Baker, & Betebenner, 2002).

Trend lines of the national percentages of all students at Grades 4, 8, and 12 who performed at the proficient level or higher on the NAEP mathematics assessments in 1990 through 2000 are shown by the solid lines in Figure 1. Also shown, by dashed lines, are the linear projections of the rates of increases that would be required to reach 100% proficient or above by 2014. As can be seen, the percentage of students scoring at the proficient level or above increased during the 1990s at all three grade levels. As measured in this metric, however, the annual increases have been quite modest, averaging a little more than 1% at Grades 4 and

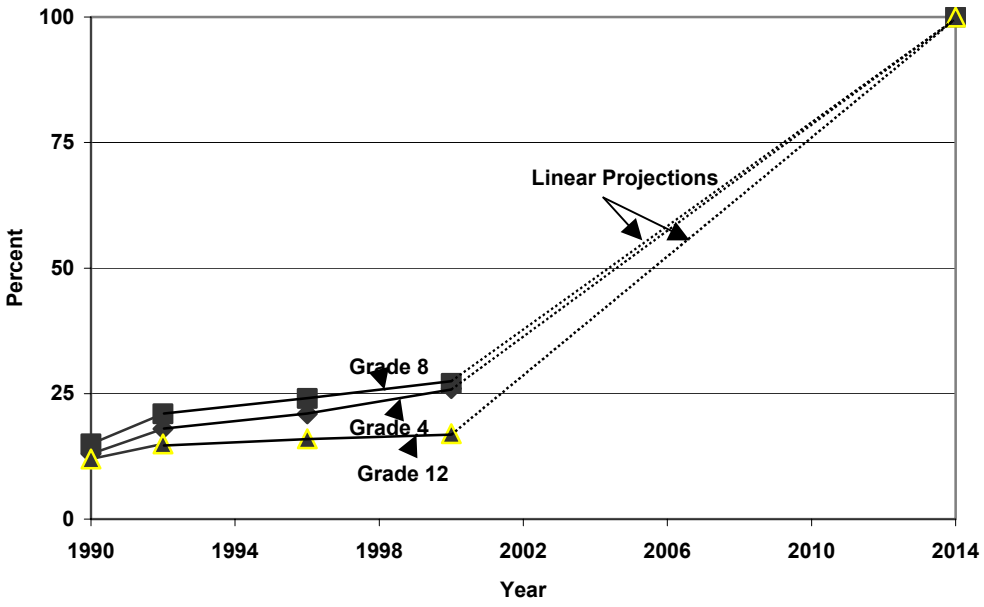

Figure 1. NAEP Mathematics trends with projections to 2014: percent proficient or above.

8, and only one half of one percent at Grade 12. Based on a straight-line projection of those rates of improvement, it would take 57 years for the percentage at Grade 4 to reach 100. For Grade 8 it would take 61 years, and for Grade 12 it would take 166 years. Looked at another way, the average annual rate of gain in percent proficient or above would have to increase by factors of 4, 4.3, and 11.8 at Grades 4, 8, and 12, respectively, to reach 100% by 2014. Such rapid acceleration would be nothing short of miraculous.

National trends in the percentages of students performing at the proficient level or above on NAEP reading assessments over the last decade show rates of increase that are less than those in mathematics. Those trends are displayed in Figure 2 along with the linear projections to reach 100% by 2014. The average increases in the percentages shown in Figure 2 are less than 1% per year. Although the percentage proficient or above figures start out at a higher level in reading than they do in mathematics, it would take even longer to reach 100% in reading than in mathematics if increases in the future are no better than they have been in the past. As was true of mathematics, a remarkable acceleration of the rate of increase in the percentage of students who are proficient or above would be required to reach the 100% goal by 2014.

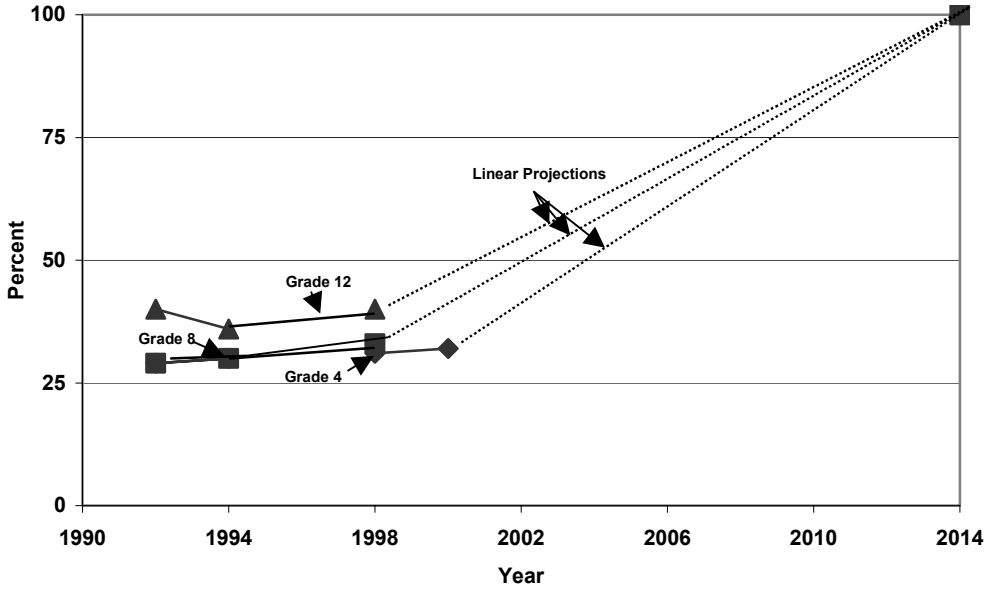

Figure 2. NAEP Reading trends with projections to 2014: percent proficient or above.

None of the states that participated in state administrations of Grade 4 reading, Grade 4 mathematics, or Grade 8 mathematics state NAEP assessments in more than 1 year showed increases in the percentage of students at the proficient level or above that were rapid enough to project anything close to 100% in the next 20 years, much less by 2014. Over the 8 years between the 1992 and 2000 state NAEP assessments of mathematics at Grade 4, the two states with the largest increases in the percentage proficient or above were Indiana and North Carolina. The increase in percentage proficient or above was 15% for both states (from 16% to 31% for Indiana, and from 13% to 28% for North Carolina). If those annual rates of increase were continued through 2014, neither state would yet be up to 65% proficient or above. North Carolina also showed the largest increase in the percentage proficient or above between the 1990 and 2000 NAEP assessments of mathematics at Grade 8 (from 9% to 30%). A continuation of that rate of annual increase through 2014 would bring the percentage up almost to 60, but still a long way from 100.

The picture is slightly better for the Grade 4 state NAEP reading assessments, not because states made more rapid gains but because the percentage proficient or above is higher in reading than in mathematics at Grade 4. Still, Connecticut, the state that had the largest increase in percentage proficient or above between the 1992 and 1998 assessments, and had a relatively high percentage to start with, had less than half (46%) of its students scoring at that level in 1998 and would need to nearly double its rate of annual increase to approach 100% by 2014.

As was noted above when discussing McLaughlin and Bandeira de Mello's (2002) comparisons of state assessments with NAEP, the performance standards set by states are generally not quite as stringent as the NAEP standards. Quite a few states have set standards that are quite ambitious, however. Indeed the idea that all students will reach the high standards that many states have set for proficient performance levels would require extraordinary acceleration in the rates of improvement of state assessment results in order to be achieved by all of their students by 2014. But, that is what NCLB regulations would require.

There are limits on the power of the federal government to bring about the educational reforms that are mandated by federal law. The federal government provides only a small fraction of the money spent on K-12 education. There is also a long tradition of state and local control of and responsibility for education. Both the distribution of relative expenditures and the tradition of control make it hard for the federal government to enforce strict compliance with federal education laws.

Experience of the Department of Education during the Clinton administration regarding compliance with the 1994 reauthorization of ESEA (Improving America's Schools Act [IASA] of 1994, Public Law 103-382) certainly would seem to support the notion of limited enforcement power. Although IASA charted a new direction for testing and reporting for purposes of Title I by the states, only 19 states were in full compliance by the time that ESEA was again reauthorized, and President Bush signed the No Child Left Behind Act in January 2002 (Robelen, 2002).

Nonetheless, Cohen (2002) has argued, "federal legislation can move states quite far, even if their actions don't all comply with the letter of the law" (p. 44). He supported this conclusion by noting the growth in the number of states that adopted standards-based reforms following the enactment of IASA. One could also point to the increase in use by states of tests that included constructed response items and were more closely tailored to state content standards as the result of IASA.

Cohen (2002) noted that Goals 2000 (Public Law 103-227) and IASA "provided a broad and flexible framework for state action, and deliberately placed considerable trust in states to work out the details for themselves. In contrast, the No Child Left Behind Act reflects significant impatience in Washington with the pace of state-led improvement and, in particular, with the slow pace at which states have instituted tough accountability systems" (p. 43). This impatience is evident in the radical change from the IASA approach of "trusting the states to work out the details" to a one-size-fits-all set of mandates regarding the grades and subjects to be tested, the reporting of results, the definition of starting points, and adequate yearly progress (AYP) objectives. The impatience also is evident in how rapidly the Department of Education moved to put out regulations for the NCLB Act and the fast-track schedule that was set for states to submit plans for meeting the NCLB requirements by January 31, 2003. The fact that all 50 states, the District of Columbia, and Puerto Rico met the January 31 deadline suggests that states have taken seriously the stricter stance of the Department of Education.

The demonstration of AYP by states, school districts, and schools is a key component of the accountability requirements in NCLB. States are required to define AYP for the state, school districts, and schools in a way that enables all students to meet the state's student achievement standards. Although the law has a number of specific constraints regarding the definitions of AYP by states, there also is some flexibility in the ways in which states define AYP.

Some Key Criteria for Definitions of AYP

The AYP definition must apply "the same high standards of academic achievement to all public elementary school and secondary school students in the State; [be] statistically valid and reliable; [and result] in continuous and substantial academic improvement for all students" (P.L. 107-110, Sec. 1111 (a)(2)(C)). Furthermore, the AYP definition must include continuous and substantial improvement in both mathematics and reading/language arts, not just for the total group of students considered as a whole, but for each of the following specific subgroups: students who are economically disadvantaged, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency.

Disaggregation of Results

Separate reporting of results for the groups identified in NCLB is laudable for several reasons. It is well known that a disproportionate number of students from the categories of students specifically targeted by the law have lagged behind on achievement tests for many years. Disaggregated reporting for those categories of students provides a mechanism for monitoring the degree to which the goal of leaving no child behind is being achieved. It also is necessary for monitoring the degree to which the achievement gap is being closed.

I believe that disaggregated reporting is desirable. It is crucial, however, that the disaggregated reporting be done in a way that yields statistically dependable information and does not identify a large number of schools due simply to the instability of school-level results when there are few students in a group for which disaggregated results are reported. For statewide reports and reports for large districts, the number of students within even the smallest category in almost all cases will be sufficient to yield results with good statistical reliability. For small districts and individual schools, however, the number of students in the less populated categories can become quite small. Hence, there is a need to consider the minimum number of students in a category that will produce results with sufficient statistical reliability to justify reporting.

States need to identify some minimum number of students required for disaggregated reporting of results. The challenge is to set a standard that will yield results that are judged to have sufficient statistical reliability to warrant reporting results that will be used to hold schools accountable. Thus, the goal in establishing a

minimum number of students is analogous to what is sometimes referred to as the "Goldilocks" standard. The minimum number should not be set so high that the potential benefits of disaggregated reporting are lost, but neither should it be set so low that there is an unacceptably high probability that schools will receive sanctions as the result of random fluctuations due to the low statistical reliability of the results for students in low-frequency categories. Unfortunately, due to the conjunctive application of the AYP requirements, schools and districts have many ways that they can fail to meet their AYP objectives.

Conjunctive Requirements

Schools can fail to meet their AYP objectives because they fall short in either reading/language arts or mathematics for any one of the subgroups of students for which disaggregated results must be reported. They can also fail to meet their objectives because test results are available for less than 95% of the students eligible to be tested. There are so-called "safe harbor" provisions that will provide relief for a few schools. A school that falls short of the AYP target for a subgroup of students will avoid being identified for improvement under the safe harbor provision if (a) the percentage of students who score below the proficient level has decreased by at least 10% from the year before, and (b) there is improvement for the subgroup on other indicators. Due to the multiple ways that a school can fail to meet AYP objectives, however, a large number of schools will be identified for improvement even with the safe harbor provision. Indeed, preliminary analyses in a number of states suggest that more than half the schools may be so identified.

Although it is reasonable to believe that all schools need to be continuously improving, the placement of a school in the needs improvement category has serious implications. First, it requires that schools develop improvement plans and districts provide public school choice. For schools identified 2 years in a row, districts need to make available tutoring services for low-income students. For schools identified 3, 4, and 5 years in a row, districts are required, respectively, to take corrective action, to develop plans to restructure schools, and to implement restructuring plans. The severe sanctions of NCLB for schools that continue to fall into the improvement category may actually hinder educational excellence because they implicitly encourage states to water down their content and performance standards in order to reduce the risk of sanctions for their schools. Furthermore, research has shown that merely reporting test results in the newspaper creates high stakes (e.g., Shepard & Dougherty, 1991), so it is possible to leverage change without such severe sanctions.

Because of the serious sanctions specified in NCLB there clearly are good reasons to avoid the over-identification of schools for improvement. There are a number of factors for states to consider in the establishment of plans so that the number of schools in the improvement category is both reasonable and manageable given the resources of districts and the state. In particular, the specifics of state definitions of performance standards, the ways in which AYP is assessed, and the specification of AYP targets for intermediate years between 2002 and 2014 all can influence the number of schools placed into the improvement category.

State Timelines

States need to specify a timeline for AYP such that all students in each subgroup listed above achieve at the proficient level or above on the state's measures of academic achievement in mathematics and in reading/language arts no later the 2013-2014 school year. Based on achievement data from the 2001-2002 school year, states are required to define a starting point for AYP. The starting point is set to be equal to the higher of the following two values: (a) the percentage of students in the lowest scoring subgroup who achieve at the proficient level or above and (b) "the school at the 20th percentile in the State, based on enrollment, among all schools ranked by the percentage of students at the proficient level" (P.L. 107-110, Sec. 111 (b)(2)(E)(ii)). If students with disabilities were the lowest scoring subgroup in the state, with 40% proficient or above, then that percentage would be compared with the percentage for the school that when ordered by percent proficient or above in mathematics was at the 20th percentile for the state based on enrollment. If 35% of the students at that school scored at the proficient level or above, then 40% would be the starting point for the state for the mathematics assessment.

What the starting point for a state will be, and therefore what the increase in percentage of students who are proficient or above will need to be between now and 2014, obviously depends on a number of factors, such as the general achievement of students in the state, the subject area, grade levels (elementary, middle, or high school), and the state demographics. The most significant factor, however, is likely to be the stringency of the state's performance standards. States with stringent standards may have starting points in the range of 20% to 40% proficient or above, whereas states with less stringent standards will have considerably higher starting points.

Approved State Plans

Five states—Colorado, Indiana, Massachusetts, New York, and Ohio—received early approval of their NCLB plans in January 2003.[2] Three of those states, Indiana, New York, and Ohio, specified the procedures that would be used to calculate their starting points, but the identification of the actual numerical values was delayed pending additional data and/or calculations. The other two states, Colorado and Massachusetts, had completed their calculations and presented the numerical starting points in their plans. The starting percentages for Colorado and Massachusetts are widely discrepant.

As shown in Table 2, the starting percentages for Massachusetts are 39.7% proficient or advanced for reading/language arts and 19.5% for mathematics (Massachusetts Department of Education, 2003). The Colorado starting percentages vary for elementary, middle, and high school, but at all three levels they are markedly higher than the Massachusetts starting percentages. Depending on the three sets of grade levels, the Colorado starting percentages for reading are roughly double the Massachusetts percentages, ranging from 74.6% to 80.3%. For mathematics, the Colorado starting percentages are as much as four times as large as the Massachusetts percentages, with a range from 50.5% for high schools to 79.5% for elementary schools (Colorado Department of Education, 2003).

Colorado and Massachusetts both participated most recently in state NAEP in 1996 for mathematics and in 1998 for reading. Table 3 shows the percentages of students in Colorado and in Massachusetts who scored at the proficient level or

Table 2

Proposed Starting Percentages Proficient or Above for NCLB on the Colorado and Massachusetts State Assessments

	Reading		Mathematics	
Grade level	Colorado	Massachusetts	Colorado	Massachusetts
Elementary school	77.5	39.7	79.5	19.5
Middle school	74.6	39.7	60.7	19.5
High school	80.3	39.7	50.5	19.5

Sources: Colorado Department of Education (2003) and Massachusetts Department of Education (2003).

[2]NCLB plans for all states, the District of Columbia, and Puerto Rico are available at the Council of Chief State School Officers Web site, www.ccsso.org/federal_programs/NCLB/1935.cfm

Table 3

Percentages of Students Performing at the Proficient Level or Above in Colorado and Massachusetts on State NAEP in Most Recent Years Both States Participated

	1998 Reading		1996 Mathematics	
Grade	Colorado	Massachusetts	Colorado	Massachusetts
4	34	37	22	24
8	30	36	25	28

Source: http://nces.ed.gov/nationsreportcard/states/

above in mathematics in 1996 and in reading in 1998. As can be seen in Table 3, the percentage proficient or above in mathematics was 2% greater in Massachusetts than in Colorado at Grade 4 (24% vs. 22%) and 3% higher at Grade 8 (28% vs. 25%). Similarly, a larger percentage of students in Massachusetts performed at the proficient level or above at both grade levels in the 1998 reading assessments (37% vs. 34% at Grade 4 and 36% vs. 30% at Grade 8). Thus, one might reasonably have expected the NCLB percentage proficient or above starting point, if anything, to be slightly lower for Colorado than for Massachusetts. At the very least, the much higher starting percentages in Colorado, compared with Massachusetts, are surprising in light of the most recent NAEP results for the two states.

The apparent discrepancy between what would be expected based on NAEP and the percent proficient or above figures for Colorado and Massachusetts has a simple explanation. The proficient level on the Massachusetts Comprehensive Assessment System (MCAS) was selected for use to determine the Massachusetts starting point for their NCLB plan. On the other hand, Colorado collapsed the four levels used for reporting the Colorado Student Assessment Program (CSAP) results to schools, districts, and the state into three levels for purposes of NCLB. The CSAP unsatisfactory level is called basic, the partially proficient and proficient levels are collapsed into a single level called proficient, and the advanced level continues to be called advanced for purposes of NCLB (Colorado Department of Education, 2003).

The Colorado plan provides the following explanation for using three levels for one purpose and four levels for another. "Colorado has defined three levels of student achievement: basic, proficient and advanced. Colorado's CSAP has four instructional levels designed to give greater detail to school personnel to better align

the state academic content standards to instruction at the classroom level" (Colorado Department of Education, 2003, p. 19).[3]

I have gone into some detail regarding the difference in the definitions of proficient in Colorado and Massachusetts not to criticize the approach taken by either state. The Colorado decision to lower the definition of proficient for the purposes of NCLB to the level called partially proficient for other purposes is, as I have argued elsewhere (Linn et al., 2002), a reasonable response to the federal legislation that would have states set a goal to be achieved by 2014 that is unrealistic. It will still be a substantial challenge to bring all students within 12 years up to the level of performance at what was called partially proficient in the past and will still be used for Colorado's own accountability system. The Colorado proficient standard for purposes of NCLB is also reasonably similar in effect to the standard set in some other states such as Louisiana and Texas.

Colorado is the only one of the five states with plans that were approved in January 2003 that combined a below-proficient category with its proficient category for purposes of NCLB reporting. The other four states used different strategies that they believe will be workable ways of defining AYP so as to have a reasonable chance of avoiding in the short run, at least, the placement of the preponderance of their schools in the needs improvement category and thereby being threatened by sanctions.

The Massachusetts and New York plans do not rely simply on the percentage of students who perform at the proficient level or above to calculate AYP. Rather, they will use index scores. The indices are defined so that they will equal 100 only when 100% of students are at the proficient level or above. But the indices also give partial credit to students who perform below the proficient level.

As is shown in Table 4, only students in the proficient or advanced categories in Massachusetts will be given full credit of 100 on the index score. Partial credit will be given, however, for students performing in the needs improvement category or even in the upper half of the warning category on the MCAS. The index score calculations are illustrated in Table 5 for a hypothetical school that has only 15% of its students at the proficient level or above. Since a substantial fraction of the students in the illustration scored in the ranges where partial credit is earned, the

[3]Colorado must submit its proficiency levels to the U.S. Department of Education's Standards and Assessment Office in Title I for final approval.

school would receive an index score of 50 rather than the score of 15 based on the percent proficient or above.

Table 4

Massachusetts Proficiency Index Scores

Performance level		Index points
Advanced		100
Proficient	Target for all students	100
Needs Improvement (High)		75
Needs Improvement (Low)		50
Failing/Warning (High)		25
Failing/Warning (Low)		0

Source: www.doe.mass.edu/ata/sprp/cycle2.pps

Table 5

Illustration of Percent Proficient or Above and Index Scores: Hypothetical Massachusetts School

Performance level	Proportion of students	Points	Index points
Advanced	.05	100	5.0
Proficient	.10	100	10.0
Needs Improvement (High)	.20	75	15.0
Needs Improvement (Low)	.25	50	12.5
Failing/Warning (High)	.30	25	7.5
Failing/Warning (Low)	.10	0	0.0
Percent proficient or above = 15			
Index score = 50			

Source: www.doe.mass.edu/ata/sprp/cycle2.pps

I should emphasize that I am not critical of Massachusetts or New York for proposing to use index scores for purposes of NCLB. On the contrary, I have previously suggested that states consider the use of index scores rather than the percentage of students who score at the proficient level or above (Linn et al., 2002). Index scores have the advantage that they give credit for increases in performance at points other than the proficient cut score. Hence, I was pleased to see that the Massachusetts and New York uses of index score were found to be acceptable by the Department of Education.

The Ohio plan (State of Ohio, 2003) takes yet another approach to establishing AYP targets that are expected to keep the number of schools falling into the needs improvement category to a reasonable number over the first few years of NCLB operation. Instead of following the linear growth model presented by the U.S. Department of Education to explain AYP, the Ohio plan specifies an AYP growth function that has increments that occur every three years until 2010, after which increments are required every year (see Figure 3).

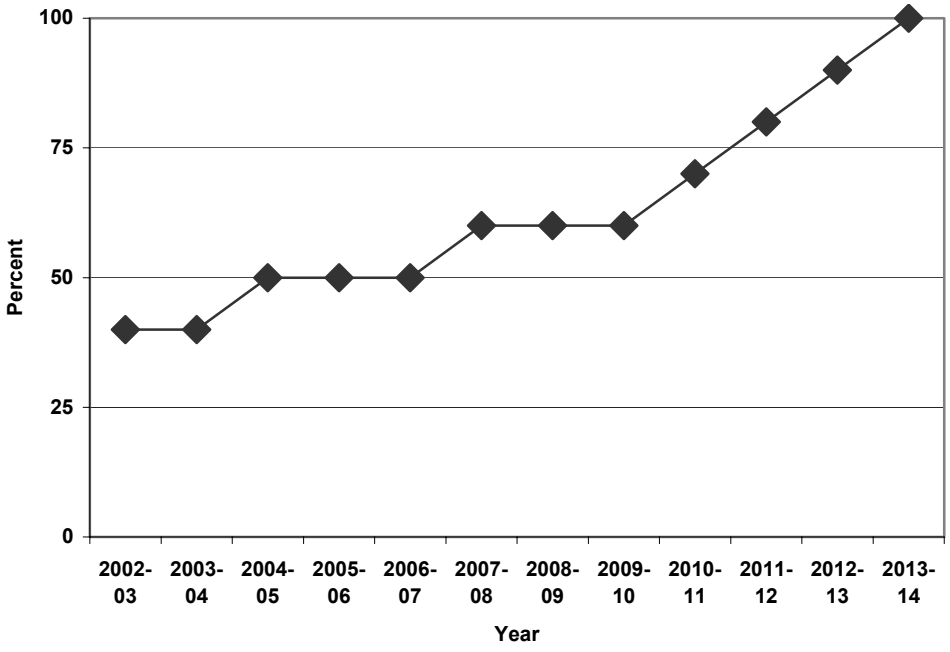

Figure 3. Ohio AYP objectives assuming a starting point of 40% proficient or above.

The rationale provided for spacing AYP increments farther apart in the early years than in the later years is that reforms take time to put in place, and therefore improvement is to be expected to take longer in the early years after the reform has been in place. Although there is a good deal of evidence from past test-based accountability systems that the gains are actually larger in the first few years than in subsequent years, the Ohio plan has considerable appeal on pragmatic grounds. A huge challenge of NCLB is for states to get through the first few years without placing an overwhelming number of schools in the needs improvement category. Buying time allows for the possibility that the law will be modified to make progress targets more realistically achievable. It is important for states to find ways to buy time with their plans for meeting NCLB requirements, and the Ohio plan is, in my view, a creative way of doing that.

Conclusion

For the last decade, or longer, accountability has held center stage for those who have shaped educational polices at both the state and federal levels. Although the focus for accountability has been largely on educators and students, the concept applies to all of us. True accountability means broadly shared responsibility, not only among educators and students, but also administrators, policymakers, parents, and educational researchers. That is, accountability means we all share responsibility for improving education. Further, if we are to meet goals to support our society's future success, this means improving opportunities at all levels of the system—pre-school, K-12, higher education, life-long learning—and effectively reaching all segments of the population.

Accountability systems have the potential to contribute to improvements in the quality of education. To do so, they need to be designed in ways that are consistent with past research evidence and experience. We know a good deal about features of accountability systems that can have positive influences on education, as well as features that have been found to be counterproductive in the past. Accountability systems need to broaden their definitions of what counts as evidence of success. Although ambitious goals are desirable, they should also be realistically grounded in past experience.

NCLB includes much that is positive. The emphasis on achievement for all students and special attention given to groups of students who have had the lowest achievement in the past are especially worthy of praise. The high aspirations are also

praiseworthy. The goals that NCLB sets for student achievement would be wonderful if they could be reached, but unfortunately they are quite unrealistic, so much so, that they are apt to do more to demoralize educators than to inspire them. If the AYP requirements are enforced, they will also result in many schools receiving sanctions that are making great strides in teaching students. This is so because steady and significant progress is not recognized as improvement under NCLB unless AYP targets are met.

Consider, for example, schools A and B that are depicted in Figure 4 along with AYP targets in a state with a 40% proficient or above starting point and straight-line AYP targets to 100% proficient or above in 2014. School A has steady and substantial increases in the percentage of students who perform at the proficient level or above. Because it starts out so far below the AYP starting point for the state, however, school A would fail to meet AYP targets each of the first 6 years of the program even without considering the performance of subgroups of students, by which time it would have been restructured. School B, on the other hand shows a quite modest increase of 1% per year in percent proficient or above—one-sixth the rate of increase

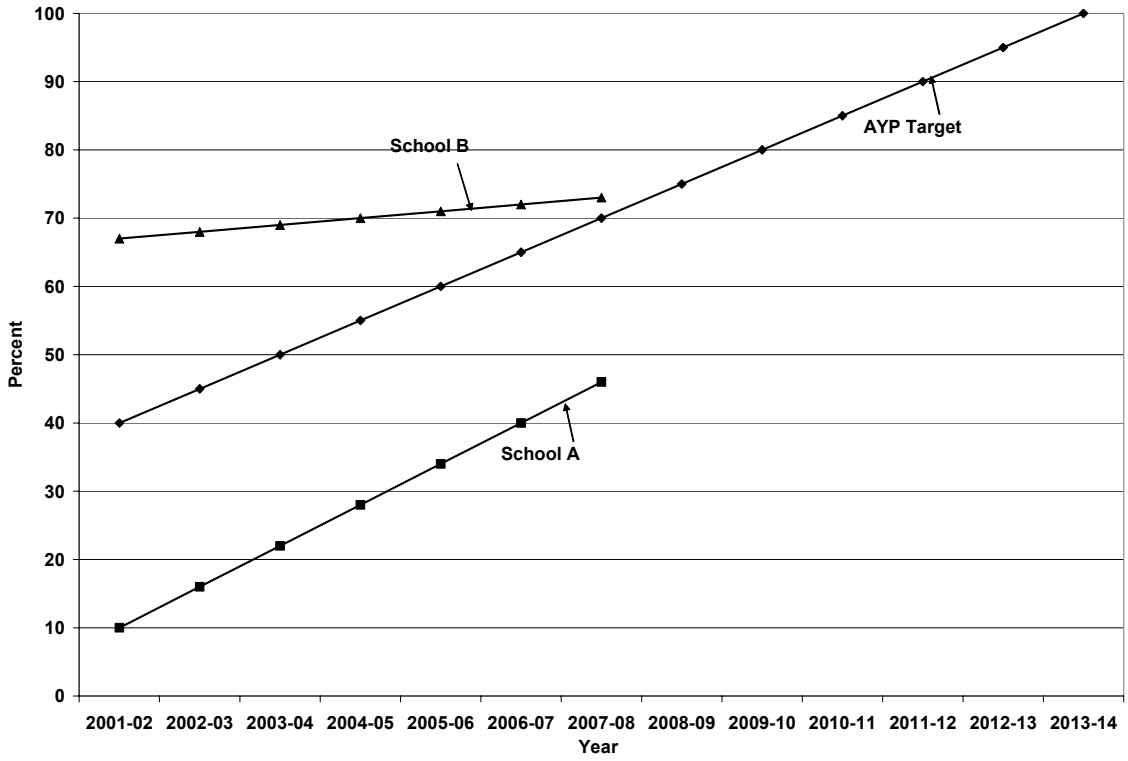

Figure 4. Progress of two schools compared to AYP target.

experienced by school A—but because it starts out well above the AYP starting point for the state, it would not be identified as in need of improvement during the first 6 years unless one of the designated subgroups of students fell below the AYP target line.

Under conditions such as those illustrated in Figure 4, it is not at all surprising that states have sought ways to minimize, in the short run, at least, the number of schools that are placed in the needs improvement category, by making expectations more reasonable to achieve. I hope that experience over the next few years with the various approaches taken by states in response to NCLB will point the way to modifications of the law and, more importantly, contribute to improved educational quality for all students, especially those students that ESEA has always been intended to help the most.

One improvement that could be made would be to use the schools that make large gains in each of the next few years as existence proofs of ambitious, but realistic targets. Suppose, for example, that the 10% of Title I schools with the largest consistent increases in the percentages of students scoring at the proficient level or above in both reading and mathematics register increases of at least, say, 4% a year. Setting an AYP target requiring an increase of 4% a year in each subject would be a challenge; indeed, it would require major improvement for the most schools, but the challenge would not be so far out of touch with reality as the targets that were plucked out of the air and dropped into the legislation. If half the schools serving low-achieving students met such a target and the rest of those schools showed improvements half that great, the overall achievement of students that Title I was designed to help would be vastly better than it is now, and the country would have reason to celebrate the quality of education provided for the nation's neediest children.

References

Cohen, M. (2002). Unruly crew: Accountability lessons from the Clinton administration. *Education Next*, *2*(3), 42-47.

Colorado Department of Education. (2003). *Consolidated state application accountability workbook for state grants under Title IX, Part C, Section 9302 of the Elementary and Secondary Education Act (Public Law 107-110).* January 5.

Elementary and Secondary Education Act of 1965, Pub. L. No. 89-10, 79 Stat. 27 (1965).

Goals 2000: Educate America Act of 1994, Public Law 103-227, § 1 et seq., 108 Stat. 125 (1994).

Improving America's Schools Act of 1994, Public Law 103-382, § 1 et seq., 108 Stat. 35424 (1994).

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

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

Massachusetts Department of Education. (2003). *Consolidated state application accountability workbook for state grants under Title IX, Part C, Section 9302 of the Elementary and Secondary Education Act (Public Law 107-110).*

McLaughlin, D., & Bandeira de Mello, V. (2002, April). *Comparison of state elementary school mathematics achievement standards using NAEP 2000*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice*, *8*(1). 14-22.

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

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1998). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Porter, A., & Chester, M. (2001, May). *Building a high-quality assessment and accountability program: The Philadelphia example*. Paper presented at the 4th annual Brookings Conference, Brookings Institution, Washington, DC.

Robelen, E. W. (2002, April 17). States, Ed. Dept. reach accord on 1994 ESEA. *Education Week*, pp. 1, 28-29.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practice*, *9*(3). 15-22.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Shepard, L. A., & Cutts-Dougherty, K. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

State of Ohio. (2003). *Consolidated state application accountability workbook for state grants under Title IX, Part C, Section 9302 of the Elementary and Secondary Education Act (Public Law 107-110).* January 6.

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

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2001). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. Boulder, CO: University of Colorado, School of Education, Education and the Public Interest Center. http://www/education.colorado.edu/epic/index.asp