Improving the Accountability Provisions of NCLB

Robert L. Linn

Center for Research on Evaluation, Standards, and Student Testing

University of Colorado at Boulder


The work herein was partially supported under the Educational Research and Development Center Program PR/Award R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education. The findings and opinions expressed in this paper are those of the author and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.
Improving the Accountability Provisions of NCLB

The No Child Left Behind (NCLB) Act of 2001 (Public Law 107-110), which reauthorized the Elementary and Secondary Education (ESEA) Act of 1965, is the centerpiece of the current administration’s educational agenda. For the 2006 fiscal year, NCLB provides 13 billion dollars in financial assistance to states, districts, and schools with high percentages of poor children to help improve student achievement. NCLB has much that is worthy of praise. It is particularly praiseworthy for its emphasis on all children, for the special attention it gives to improving learning for children who have lagged behind in the past, and for the attention given to closing persistent gaps in achievement.

A critical, albeit less praiseworthy, feature of NCLB is the heavy reliance that the law places on test-based accountability. Although there is widespread support for the general idea that schools and educators should be held accountable for improving student achievement, some of the specific NCLB accountability requirements are unrealistic or counter productive and threaten to undermine the praiseworthy aspects of the law.

The NCLB accountability system has several fundamental problems that need to be addressed. Five of the most serious problems will be discussed in this paper and proposals for improvement will be provided. A brief summary of the five fundamental problems that need to be addressed is first presented before turning to more detailed consideration of each of the problems in turn.

Overview of Five Problems

1. *Expectations*. The expectations for student performance, which include the adequate yearly progress (AYP) targets that lead to all students performing at the proficient level or above
by the 2013-2014 school year, are unrealistic and unobtainable (Lee, 2006; Linn, 2003a). More realistic expectations need to be established that are ambitious, but obtainable given sufficient effort on the part of educators and students.

2. **Proficiency.** Proficient student achievement, while fundamental to the determination of AYP, is so poorly defined and varies so much from state to state that it has become a meaningless concept. A better definition of acceptable or desirable levels of student achievement is needed.

3. **Fixed Targets vs. Improvement.** Although the “P” in AYP stands for progress, the definition of AYP focuses on current student achievement each year in comparison to a fixed target. With the exception of the rarely used safe harbor provision, improvement in student achievement is not considered in determining the AYP status of a school. A revised NCLB accountability system should consider both improvement and status.

4. **Narrow Measures.** NCLB focuses narrowly on state assessments of mathematics and reading or English language arts. Although achievement in these subjects is obviously important and state assessments can provide relevant measures of achievement, there are other important subjects and other indicators of student achievement. An improved NCLB accountability system should allow for the use of multiple sources of information about student achievement.
5. **Multiple Hurdles.** There are many ways that a school can fail to meet the AYP requirements in a given year, but only one way that it can meet them. It must meet or exceed the participation rate requirements (95% of eligible students) for mathematics and reading/English language arts for the student body as a whole and for each subgroup of students where disaggregated reporting is required, and must meet or exceed the percent proficient or above targets for all students and for all subgroups. An improved NCLB accountability system would allow for high performance in one area to compensate to some degree for lower performance in another area.

**Unrealistic Expectations**

The NCLB goal for student achievement is that all students perform at the “proficient” level or above by the 2013-2014 school year. As is discussed in detail in the following section of this paper, proficient achievement is in reality a poorly defined concept that has come to have wildly different meanings in different states. The stated intention of NCLB, however, is that the proficient achievement standard should correspond to a high level of student achievement. Specifically, NCLB requires states to set challenging academic achievement standards that (I) are aligned with the academic content standards; (II) describe two levels of high achievement (proficient and advanced) that determine how well children are mastering the material in the State academic content standards; and (III) describe a third level of achievement (basic) to provide complete information about the progress of lower-achieving children toward mastering the proficient
and advanced levels of achievement. (NCLB, 2001, Part A, Subpart 1, Sec. 1111 (b) (D) (ii)).

Although the actual definition of academic achievement standards defining the basic, proficient, and advanced performance levels is left to the states, it is clear that the intent of the law is that the proficient and advanced levels should be set at ambitious levels. Ambitious achievement standards are in keeping with the spirit of the standards movement that has dominated educational assessment and accountability for more than a decade. Certainly, when achievement levels were set on the National Assessment of Educational Progress (NAEP) for the first time in 1990, they were set at quite high levels. When NAEP mathematics results were first reported in terms of achievement levels in 1990 a 4th grade student had to be at the 87th percentile to be counted as proficient. The minimum to be counted as proficient corresponded to the 85th percentile at the 8th grade and the 88th percentile at the 12th grade (Braswell, Lutkus, Grigg, Santapau, Tay-Lim & Johnson, 2001). Thus 15% or fewer of all students performed well enough on the 1990 NAEP mathematics assessments at any of the three grade levels assessed to be considered to be proficient or above.

Although the NAEP standard for proficient achievement was clearly set at an ambitious level in mathematics at all three grade levels, there were no consequences for students or schools of failing to meet that standard. Under NCLB, however, there are real consequences for falling short of the proficient level on state assessments. Targets for making adequate yearly progress (AYP) are established in terms of the percentage of student scoring at the proficient level or above on state assessments and those targets must increase
in a pattern that leads to 100% of the students being proficient in both reading or English language arts and mathematics by 2013-2014.

The expectation that a high level of achievement is required to reach the proficient level is in keeping with the achievement levels set on NAEP and on the assessments in a number of states prior to the enactment of NCLB. But, is it realistic to require that all students perform at such a high level in 2013-2014? When the year for achieving the goal was selected it was 12 years away from the initial year of NCLB, 2002, and thus corresponded to the 12 years of elementary and secondary education or to the idea that a 1st grade student in 2002 should reach the proficient level by the 12th grade. Even if that trajectory were reasonable for all students, which is highly debatable, it doesn’t allow many years for the student who is in the 3rd grade in 2013-2014 and may have started with an extremely low level of achievement as a 1st grade student in 2011-2012, yet the latter student is also expected to be proficient in 2013-2014.

Each state must have its own assessment and set its own proficient level on that assessment. Starting in 2003 states must also participate every other year in state administrations of NAEP in reading and mathematics at grades 4 and 8. The requirement to participate in NAEP does not specify how the state NAEP results will be used, but it is clear that NAEP is expected to provide a benchmark against which performance on individual state assessments could be compared. NAEP provides the only common measure of student achievement at grades 4 and 8 and the only uniform definition of proficient achievement across states.

One of the potential uses of the NAEP results is to evaluate the reasonableness of the expectation that all students will achieve at proficient
level or above by 2013-2014 where the high level of achievement demanded by NCLB is needed to be proficient. Some would like to have NAEP become an official benchmark for judging state standards. For example, Governor Jeb Bush and Mayor Michael Bloomberg recently suggested that NAEP “should become the official benchmark for evaluating state standards.” They went on to suggest that states should be required to bring their standards into line” with NAEP (Bush & Bloomberg, 2006 p. 3).

In the years since NAEP achievement levels were first established in mathematics in 1990 there have been fairly substantial increases in the percentage of students performing at the proficient level or above on the NAEP mathematics assessments. The gains have been most noticeable at the 4th grade, but there have also been substantial increases at the 8th grade. On the 2005 NAEP mathematics assessment 36% of students at grade 4 were at the proficient level or above compared to only 13% in 1990. At grade 8 the percentage of students proficient or above on the mathematics assessments increased from 15% in 1990 to 30% in 2005 (Perie, Grigg, & Dion, 2005). Although these increases in performance are encouraging, the rate of improvement is not nearly rapid enough to reach the goal of 100% proficient or above by 2013-2014. From 1990 to 2005 the percentage of students at the proficient level or above increased by an average of 1.53% per year at grade 4 and 1.0% per year at grade 8. If those trends continued for the 9 years from 2005 to 2014 the percentage of students that would be at the proficient level or above in 2014 would be 50% at grade 4 and 39% at grade 8. Half the students is a long way from all students.
The achievement levels for the reading assessments were set in 1992. In that year the proficient cut score corresponded to the 71st percentile at both grades 4 and 8, that is, 29% of the students were at the proficient level or above. Since that time there has very little increase in the percentage of students performing at the proficient level or higher in reading. Thirty-one percent of public school students were at the proficient level or above in 2005 at both grades 4 and 8 (Perie, Grigg, & Donahue, 2005). If that same trend of a 2% increase over 13 years were to continue for the next 13 years, there would still be only a third of the students at the proficient level or above in reading at grades 4 and 8 in 2018. Thus, 4 years after the 2014 NCLB deadline for having all students reach at least the proficient level, two thirds of the students would fail to achieve that level of performance. Radical changes clearly would be needed in the trends to even come close to the NCLB 100% goal and there is no reasonable basis for thinking that such changes are feasible.

The national trends on NAEP do not auger well for reaching the 100% proficient goal by 2014. It is simply unrealistic to expect that there would be a sudden and sustained acceleration in the percentage of students reaching the proficient level to achieve the NCLB goal of 100% (Lee, 2006; Linn, 2003a). But, what about results for individual states? Are there states that have shown much better achievement in general and in rate of improvement as measured by NAEP? As was required by NCLB, all 50 states participated in state administrations of NAEP in reading and mathematics at grades 4 and 8 in 2003 and again in 2005. Certainly there are states that have better achievement than the nation as a whole and there are states that had greater increases in student
performance than was observed for the nation. Of course, the flip side is also true, that is, students in some states performed more poorly than the nation.

Figures 1 through 4 display the 2003 to 2005 trend lines in the percentage of students at the proficient level or above for reading at grades 4 and 8 and mathematics at grades 4 and 8. In 2005 the percentage of students at the proficient level or above in reading at grade 4 (Figure 1) ranged from a low of 18% in Mississippi to a high of 44% in Massachusetts. At grade 8 the range was from 19% in Hawaii to 44% in Massachusetts (Figure 2). The corresponding ranges for mathematics were slightly larger than the ones for reading – from 19% in New Mexico to 49% in Massachusetts at grade 4 (Figure 3) and from 13% in Mississippi to 43% in Massachusetts at grade 8 (Figure 4). Clearly the states differ in both subjects and both grade levels in the percentage of students who are at the proficient level or above on NAEP, but no state has even half their students at that level in either reading or mathematics at either grade.

< Insert Figures 1 through 4 about here >

Figures 1 and 2 show, not only how far states are from the 100% goal, but that trend lines are generally flat for reading at both grade levels. The increases in percent proficient or above shown in Figures 3 and 4 were usually larger for mathematics than for reading, but only 3 states, Idaho, Massachusetts and Montana had increases that averaged 4 or more percentage points per year at grade 4 and only one state, Texas, had an increase at grade 8 that averaged as much 3 percentage points per year.

If the proficient achievement standard is set at a high level as prescribed by NCLB and reflected on NAEP, then the goal of having all students
performing at that level by 2014 is out of reach despite intensive efforts of educators and students. Unobtainable goals might not be a bad thing if there were no consequences for failing to reach them. Failure to meet NCLB goals, however, has serious consequences for schools, educators, and students. Schools that fail to meet AYP targets two years in a row are placed in the needs improvement category. Those schools must develop a school improvement plan, offer supplemental educational services such as tutoring, and offer school choice. Failure to make AYP for a third year in a row can result in corrective action such as extending the school year or replacing school staff and schools that fail to make AYP 5 years in a row are subject to restructuring. Setting unrealistically ambitious goals and sanctioning schools that fail to meet them does more to demoralize than to motivate educators. This is not an argument against sanctions or against having ambitious goals. Rather, it is an argument that goals should be ambitious, but they also should be realistically obtainable given sufficient effort, especially when there are sanctions for not reaching them.

At the very least there should be an “existence proof” that goals are obtainable. In the case of NCLB, it should be possible to identify some schools that have either already achieved or are on a clear trajectory for achieving the goals that are set for all schools. One way of assuring that goals are ambitious, but still within reach, is to use empirical results to set the goals. Schools might be rank ordered in terms of the rate of improvement in student achievement on the state’s assessments in reading or English language arts and in mathematics over the past 4 or 5 years. The highest ranking, say 20%, of schools in terms of gains made on each assessment could then be used to set the goal for all
schools. If the top 20% increased the percentage of students performing at the proficient level or above by an average of, say, 2% per year in reading/English language arts and, say, 3% per year in mathematics then increases of 2% and 3% per year could be set as the goals for reading or English language arts and mathematics respectively for all schools. Those would certainly be ambitious goals for schools that had shown little if any improvement or possibly even declined during the past 4 or 5 years, but it would also be based on the knowledge that continued improvement at the identified rate is possible as demonstrated by the performance of the set of schools that were used to set the goals.

It is recognized that the target percentage improvement for assessments in the content areas might vary in how challenging they are to achieve for schools starting out with low student achievement than for schools starting out with high achievement. The comparability of the challenge would certainly be more similar for schools starting at different levels of student achievement, however, than the current system that sets the same target each year for all schools and sets the goal to be achieved by 2013-2014 at 1000% proficient or above. The improvement targets could also be combined with an overall status target so that schools that reached some relatively high level of achievement would not be expected to continue to show substantial improvements in student achievement each year.

**Proficiency**

The standards movement has its roots in the 1980s starting with the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983) and the efforts of the National Governors Association (NGA)
that led to the 1989 Charlottesville Education Summit. President George H. W. Bush and the governors concluded that national goals were needed to stimulate reform of the educational system. President Bush announced the goals in his 1990 State of the Union Address and the National Educational Goals Panel (NEGP) was established later that year (see, for example, Vinovskis, 1999). The NEGP played a leadership role in establishing and defining the standards-based reform movement. Standards-based educational reform involves the identification of two types of standards, content standards that specify the subject matter material to be taught and performance standards that specify expected levels of knowledge and understanding of that material for students.

The 1989 Charlottesville Education Summit and the NEGP encouraged the establishment of ambitious performance standards and provided the context in which the performance standards, referred to as achievement levels, were set for NAEP. As was described above, the NAEP achievement levels defined quite high expectations. Standards-based education and accountability was a prominent feature two pieces of legislation enacted during President Clinton’s Presidency. Content standards, student performance standards, and standards-based assessments were key ideas of the Goals 2000: Educate America Act of 1994. Standards were also played an important role in the program evaluations required in the re-authorization of ESEA by the Improving America’s Schools Act (IASA) of 1994 (see, for example, Hess & Petrilli, 2006, chapter 1).

Using performance standards, or student academic achievement standards, as they are referred to by NCLB to report results of student assessments is seen as preferable to the older ways of reporting test results in
terms of averages and national norms. The achievement standards are meant to provide a way of reporting the degree to which student achievement is meeting expectations rather than merely indicating how students are performing in comparison to other students. The achievement standards are supposed to provide a means of identifying the level of performance that is considered good enough and thereby establishing a goal for student achievement.

Reporting assessment results in terms of academic achievement standards would be sensible if there was a good way of deciding the level of achievement that corresponds to proficient performance. Unfortunately, the setting of achievement standards is fraught with difficulties. There is widespread agreement in the measurement profession even among those who are most in favor of setting achievement standards that “…there is NO true standard that the application of the right method, in the right way, with enough people will find” (Zieky, 1995, p. 29). Furthermore, there is no scientific basis for defining proficient performance or translating that definition into a cutscore on an assessment. Rather, the achievement standards are set by judges using one of several different methods to translate their judgments into cutscores on an assessment. Where the cutscore gets set depends on the method used, the context in which judgments are made, and the judges that are used in the process (Glass, 1978; Jaeger, 1989; Linn, 2003b, in press).

Given the uncertainties in standard setting it is not surprising that it turns out that there is wide variability in the stringency of cutscores that states have set to define proficient achievement on their assessments. Olson (2005) published the percentage of students who were reported to be proficient or above on reading or English language arts and mathematics assessments at
grades 4 and 8 in 2005 for each of 47 states that had assessments in place at that point\textsuperscript{1}. The percentage of students who were proficient or above on the state reading/English language arts assessments ranged from a low of 35% in Missouri to a high of 89% in Mississippi at grade 4. The corresponding range at grade 8 was from 30% in South Carolina to 88% in North Carolina. The ranges of percentages proficient or above according to state mathematics assessments ranged from 39% in Maine and Wyoming to 92% in North Carolina and from 16% in Missouri to 87% in Tennessee, at grade 4 and 8, respectively.

Although as was illustrated in Figures 1 through 4 there certainly are real differences in achievement between states, the differences are not nearly as great is suggested by the variation in reported percentage of students who are proficient or above at in either subject at both grade levels., Maine, Missouri, South Carolina, and Wyoming have clearly defined proficient achievement in a much more demanding way than the definitions used by Mississippi, North Carolina, and Tennessee. It is nonsensical to think that achievement in mathematics at grade 8 in Tennessee is so much better than it is in Missouri that the percentage of students who reach the proficient level or above is more than 5 times as large in Tennessee as it is in Missouri.

State-by-state NAEP mathematics results in 2005 at grade 8 stand in sharp contrast to the state assessment results reported for Tennessee and Missouri. The NAEP percentages of students who were proficient or above for those two states differed little, but the percentage was slightly lower for Tennessee (21%) than Missouri (26%) which are dramatically different than the 87% and 16% reported for the state assessments. The definitions of proficient

\textsuperscript{1} The closest grade was used for the few states that did not have assessments in place at either grade 4 or 8 in 2005.
achievement vary so wildly from state to state that the concept lacks in semblance of a common meaning. Certainly, reporting results in terms of percent proficient on state assessments lacks comparability from state to state.

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

Another alternative would be to compare change over time to the current variability of student performance. This is usually done by dividing the change in mean scores by the standard deviation of scores, a measure often called an “effect size.” The use of effect size does not suggest any acceptance of the current distribution of performance as adequate. Rather, effect size statistics are used routinely because they make available a variety of evidence that can
used to gauge the rate of change—e.g., historical data on trends, international comparisons, and evaluations of the effects of past educational interventions.

An effect size for 2003 would be equal to the difference in the mean achievement score in 2003 and the mean in 2002 divided by the standard deviation in 2002. Using effect size statistics for the top performing, say 20%, of schools, an annual target increase in effect size could be set. This might result in an annual target increase in effect size of, say, .05. Thus, a school would need to have a cumulative effect size of .6 when 2014 was compared to the 2002 base year. An effect size of .6 would mean that average student in 2014 would be performing at a level equivalent to roughly the 73rd percentile in the 2002 distribution. That would represent a larger increase in student achievement of the nation’s students than has been achieved in any similar period of time during the last 50 years. Although effect size statistics are seemingly less transparent to the public than the percentage of students scoring above a cutscore, effect size statistics are widely used and are certainly less complicated to explain than the confidence intervals now used by most states to determine AYP.

**Fixed Targets vs. Improvement**

Although the NCLB accountability system might appear to focus on improvement as suggested by the word progress in AYP, it actually focuses on current status. Schools where students who are already achieving at relatively high levels, for example, can actually have a decline in achievement from one year to the next, and still make AYP. Schools with very low achievement initially, on the other hand, will routinely fail to meet AYP even if they show rather sizeable year-to-year gains in student achievement. This is so because,
with the exception of the rarely applicable safe harbor provision, AYP focuses on current achievement in a given year in comparison to an Annual Measurable Objective (AMO) for that year rather than changes in achievement from one year to the next. Consequently, schools that have a high achieving level to begin with have a relatively easy time meeting AYP without any gains in achievement, at least in the first few years. On the other hand, schools with initially low achieving students would have to have extraordinary improvement in achievement to meet AMO. Consequently, many schools that are actually showing considerable progress, and deserve recognition for the gains they are making, fail to meet AYP because of their initial low performance.

Basing evaluations of schools almost exclusively on current performance of students in relationship to fixed targets ignores the fact that schools differ substantially in the achievement of their students when they enter school. It privileges schools serving students who are already high achieving and puts schools serving initially low achieving students at a substantial disadvantage. The inference that a school A is of low quality or that the teachers in school A are less effective than those in School B based solely on the fact that the percentage of students who are at the proficient level or above in a given year is smaller in school A than the corresponding percentage at school B is simply not justified because there are so many other possible explanations of the difference, most notably that the students in the two schools differed in their levels of achievement at the start of the year or when they entered first grade.

Many state devised school accountability systems base their evaluations of schools on a combination of current status measures and improvement in student achievement from one year to the next. Therefore it is not surprising
that a number of states have expressed interest in the possibility of changing the way in which AYP is determined for NCLB to allow greater emphasis on improvement.

A change in the NCLB accountability system that would allow schools to meet AYP either because their current achievement met a target or because the improvement in achievement met an improvement target seems desirable. This might be accomplished with a less stringent safe harbor criterion. Consistent with proposals above, both the current year achievement target and the improvement target should be set in light of what has been shown to be possible by schools that have shown substantial gains over a period of 4 or 5 years.

An alternative way of evaluating change in achievement that is attractive to several states is the use of longitudinal student records to track the growth in achievement for individual students. Analytical procedures, commonly referred to as value-added models, are used to estimate the school effects on student growth. Consideration should be given to the possibility of allowing states to use results of value-added analyses to provide evidence of improved achievement. The value-added results could be used, possibly in combination with status measures, to satisfy AYP requirements.

In response to widespread interest in approaches that focus on growth for purposes of determining AYP, the U.S. Department of Education authorized a pilot program that allowed states to submit proposals to use a growth model to make AYP determinations. The pilot program was announced by Secretary Spellings on November 21, 2005. Several “core principles” that must be met for a proposal to be approved were identified in a letter from Secretary Spellings to
the Chief State School Officers regarding the pilot program. The first, and perhaps the most constraining principle, specifies that the growth model “must ensure that all students are proficient by 2013-2014 and set annual goals to ensure that the achievement gap is closing for all subgroups of students” (Spellings, 2005). Thus, despite the argument that the expectation is unrealistic, the fixed achievement target of 100% proficient or above in 2013-2014 is maintained.

Eight states submitted proposals to participate in the growth model pilot program and two of those proposals (North Carolina and Tennessee) were approved for implementation of growth model pilots in 2005-2006 (Spellings, 2006). Early consideration of revised proposals for possible implementation in 2006-2007 is expected for the other six states (Alaska, Arkansas, Arizona, Delaware, Florida, and Oregon).

The pilot program takes one step toward a system that would use information about improvement in student as well as current achievement in determining whether or not schools are performing adequately. This is an important step, but so far will be applicable only for a small fraction of the states. It is also limited by the continuing requirement that the amount of growth will lead to all students reaching at least the proficient level by 2014. The option of using improvement as well as current status to determine AYP needs to be available to more states and, as was argued above, more realistic achievement goals need to be set.

Many states lack a longitudinal data system that would allow them to implement a value-added model. Improvement in performance of students in those states could still be used in the determination of AYP by comparing the
performance of student cohorts from one year to the next. Comparisons of successive cohorts of students (e.g., 4th grade student in 2006 compared to 4th grade student in 2005) lacks some of the advantages of longitudinal tracking of student achievement, but can still provide information on changes in student achievement that would complement the comparisons of current performance to fixed targets each year.

**Narrow Measures**

The NCLB school accountability system uses student achievement results on state assessments in just two subjects, mathematics and reading or English language arts, to determine AYP. The ability to read and achievement in mathematics are obviously of critical importance. But these are not the only subjects that are vital for students to learn. Other subjects (e.g., science, history, and civics) are also important for students to understand to be prepared to function in modern society.

State assessments are not the only means of measuring student achievement. Student achievement in reading, mathematics, and other subjects is routinely measured in a variety of ways, including district benchmark assessments, formative classroom assessments, and teacher grades. These other indicators can provide relevant information about student achievement not only in the two subject areas that are privileged by NCLB, but in other content areas as well.

In a summary report of public hearings on NCLB the Public Education Network (2006) reported that the public believes that the focus of the NCLB accountability system is too narrow. Concern was expressed that the results often conflict with the results of a state’s own accountability system and with the
personal evaluations of schools by members of the public. The public thinks that there is too much emphasis on a single assessment as the determining factor for AYP and would like to see the emphasis reduced by making greater use of formative evaluation information.

The public has good reasons to be concerned about the narrow focus on state assessment results in mathematics and reading/English language arts for the determination of AYP. A survey of school districts conducted by the Center on Education Policy (CEP) (2006) found that 71% of the districts said that they had reduced time devoted to at least one other subject to allow more time to be devoted to reading and mathematics (p.89). Although some of the districts indicated that the additional time spent on reading and mathematics had helped low achieving students make gains in those subjects “others reported that students were shortchanged in important subjects like social studies and science” (CEP, 2006, p. 89).

The CEP finding that less time was devoted to on non-tested subjects is consistent with results reported in other studies. A substantial majority of teachers in two districts surveyed by Sunderman, Tracey, Kim, & Orfield (2004), for example, reported that AYP requirements caused some teachers to increase the amount of time spent on activities specifically aimed at preparing students for state-mandated assessments while de-emphasizing or neglecting content of untested topics.

The nearly exclusive emphasis on state assessments of reading and mathematics can not only narrow the curriculum by overemphasizing those subjects at the expense of other subjects, but it can also narrow the teaching of the two target subjects by limiting instruction to the material that appears on the
assessments. Narrow teaching to the test emphasizing formats and predictable patterns of questions can lead to inflated test scores. Koretz (2005) has defined score inflation “as a gain in scores that substantially overstates the improvement in learning it implies” (p. 99). Score inflation is not unusual in high-stakes accountability uses of assessment results. Several studies have found that gains in scores on high-stakes tests do not generalize well to other indicators of student achievement such as results on the National Assessment of Educational Progress (NAEP) or the ACT (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar and Shepard, 1991; Linn, Graue, & Sanders, 1990).

Koretz (2005) has noted that the measurement community has frequently spoken out against the use of a single test score for making important decisions largely because errors of measurement, that are an unavoidable part of any test, may lead to incorrect decisions. The use of multiple measures can mitigate problems associated with measurement error. Koretz (2003, 2005) also noted that there is an additional reason for favoring the use of multiple measures in high-stakes, test-based accountability: namely that a single assessment can intensify score inflation by exacerbating undesirable incentives to narrowly teach to the test, to say nothing of the occasional educator who will succumb to the temptation to cheat.

The state accountability system that has been in place in Kentucky for more than a decade provides a good illustration of a school-building accountability system that encourages the teaching of a broad array of academic subjects. Student assessments in seven content areas (reading, writing, mathematics, science, social studies, arts and humanities, and practical
living/vocational studies) are used by the Kentucky accountability system in judging the academic progress of a school (http://www.kde.state.ky.us/).

Although the heaviest weight is given to reading and mathematics at the elementary school level, science, social studies, and writing also are given substantial weights. For middle schools and high schools reading, mathematics, science, social studies, and writing are given equal weights while somewhat lower weights are assigned to arts and humanities, and practical living/vocational studies. The Kentucky approach to school accountability clearly encourages educators to attend to more than reading and mathematics.

Although college admissions decisions differ in many ways from school accountability decisions, there are lessons to be learned from experience with college admissions that may have relevance for school accountability. There is a vast literature on the use of standardized tests to predict performance in college (see for example, Zwick, 2002). The research is quite consistent in showing that college admissions tests such as the SAT and ACT provide reasonably good prediction of student performance in college as measured by grade-point average. In most instances, however, high school grades or rank in class provide better prediction of college grades than is provided by standardized tests. This is so, despite the fact that there are obvious between-school differences and between-course differences in the stringency of high school grades. Moreover, the combination of high school grades or rank in class and the standardized test scores do a better job of prediction than either alone.

The results of studies that have investigated the prediction of performance in college suggest that teacher assigned grades contain useful
information about student achievement despite the fact that the grades are not
standardized. The use of teacher assigned grades or systematic ratings of
student achievement provided by teachers, together with a common set of
district-selected benchmark assessments, and formative classroom
assessments selected by teachers, would likely improve the quality of the
information about student achievement and could increase the validity of school
evaluations when combined with the results of state assessments.

Teacher ratings of student achievement in selected subjects or teacher
selected classroom assessments could broaden the sources of information
about student achievement. Combining teacher produced scores with state
assessments would require that the teacher scores be reported in a common
metric such as a 1 to 5 scale and concerns that teachers might report inflated
ratings would need to be addressed, but the potential gain in information would
be worth the added effort needed to obtain and use teacher produced scores.

Multiple Hurdles

A school can fail to make AYP for many different reasons. At a minimum
a school must clear five distinct hurdles to make AYP in a given year. A school
must have at least 95% its eligible students participate in the mathematics and
in the reading or English language arts assessments. The percentage of
students who perform at the proficient level or above must meet or exceed the
AMO for mathematics and the AMO for reading/English language arts. It must
also meet the criterion set for the additional academic indicator selected by the
state (e.g., attendance rate for elementary and middle schools and graduation
rate for high schools). Because of disaggregated reporting requirements for
subgroups, schools with diverse student bodies are frequently confronted with
many more than the 5 hurdles based on all students in the school. As the number of subgroups for which disaggregated reporting is required increases, the number of hurdles that a school must clear rapidly increases (Marion, White, Carlson, Erpenbach, Rabinowitz, & Sheinker, 2002).

Thus a school with more than the minimum number of students in each of several subgroups identified for disaggregated reporting has substantially more than 5 hurdles to clear. For example, a school with 6 subgroups (African American students, Hispanic students, white students, students with limited English proficiency, economically disadvantaged students, and students with disabilities) meeting the minimum size requirement would have not 5, but 29, hurdles to clear (the 5 when all students in the school are considered as a whole, plus 24 for the 4 hurdles (participation rates in reading and mathematics, and achievement in reading and mathematics), for each of the 6 subgroups. Thus, the latter school could fail to make AYP in 29 different ways but could make AYP in only one way – by clearing all 29 hurdles.

Requiring schools to meet AYP requirements for separate subgroups of students is consistent with the NCLB goal of closing gaps in achievement for the identified subgroups. Nevertheless, it is clear that NCLB’s multiple-hurdle approach makes it considerably more difficult for large schools with diverse student bodies to meet AYP requirements than it is for small schools or schools with homogenous student bodies (Kim & Sunderman, 2005; Linn, 2005).

There are alternatives to the conjunctive system of multiple hurdles used in the NCLB school accountability system. The most obvious alternative is some form of a compensatory system. With a compensatory approach, high achievement that is above the goal in one content area can be used to
compensate for achievement that falls below the goal in another area. If the AMO for a given year was 50% proficient or above in reading and 40% proficient or above in mathematics, for example, then a school where, say, 55% of its students were proficient or above in reading but only 38% of its students were proficient or above in mathematics could make AYP under a compensatory system while it would fail to do so under the current multiple-hurdle system. A number of state accountability systems that were in place prior to the enactment of NCLB used a compensatory approach. The California and Kentucky state accountability systems provide illustrations of compensatory approaches.

California’s accountability is based on a compensatory approach. The California Academic Performance Index (API) is a weighted combination of scores on tests of English-language arts, mathematics, history-social science, and science as well results of the California Alternate Performance Assessment. Schools are evaluated based on their overall API results rather than separately by comparing results to fixed targets subject by subject. The Kentucky state accountability system also uses a compensatory approach. An academic index is calculated for the Kentucky accountability system as a weighted composite of assessment results at selected grades in the seven content areas assessed by the state. An overall accountability index is then computed by combining the academic index with non-academic indicators such as attendance and retention that are assigned modest weights. Relatively low achievement in one or two content areas can be compensated by achievement that exceeds expectations in other content areas.
A conjunctive, multiple-hurdle approach such as that used to determine AYP has at least two disadvantages. First, as was already noted it places large schools with diverse student bodies at a disadvantage in comparison to a school that may serve students that are equally at risk but are more homogeneous demographically. Second, overall reliability of a school’s classification can be no greater than the reliability of the least reliable indicator. A compensatory system can ameliorate both of these disadvantages. The first disadvantage alleviated by reducing the number of indicators and associated targets. The second disadvantage is mitigated because the composite measure is more reliable than the individual measures that make up the composite.

Hybrid systems are also possible and in the case of NCLB might be preferable to either a pure compensatory system or a pure conjunctive system with many hurdles. A composite index could replace the five separate hurdles of participation rates and performance in the two subjects and on the other indicator. Composite indices for subgroups could also be created for the subgroups for which disaggregated reporting is needed based on a combination of participation rates and performance in reading/English language arts and mathematics. The multiple-hurdle feature of the hybrid system would then be the requirement that targets be met for subgroups as well as for all students in a school. Such a system would retain the focus on subgroup performance and closing of achievement gaps while allowing compensation across subject areas. It would also be a way to add the science assessments that will be required in 2007-2008 into the mix for determining AYP without adding to the number of hurdles that a school must clear. Moreover, it would make it relatively easy to
add other measures of achievement obtained from district benchmark assessments and formative assessments selected by teachers. An overall achievement index could be formed as a weighted combination of all the measures included in the system.

**Conclusion**

NCLB has the potential to make substantial positive contributions to education. It can contribute to the improvement of student achievement and, through its focus on students who have lagged behind and too often been ignored in the past, to the closing of achievement gaps among racial/ethnic groups, between economically disadvantaged students and their more affluent counterparts, between limited English proficient student and native English speakers, and between students with and without disabilities. Some features of the NCLB accountability system, however, need to be modified if the praiseworthy goals of NCLB are going to be achieved.

The most important modification is to set performance targets for judging adequate yearly progress that are more reasonable and for which there is a realistic hope that they might be achieved given sufficient effort. The need for more realistic goals applies to both the safe harbor provision of the law and to the annual performance targets. The current definitions of proficient achievement established by states lack any semblance of a common meaning. Alternatives to defining proficiency should be considered that would provide more meaningful and comparable achievement targets. Past data on what schools showing exemplary gains in achievement should be used to set goals that are ambitious, but obtainable with hard work and those goals should be
expressed in ways other than the currently poorly defined proficient academic achievement standards that vary wildly from state to state.

Changes to AYP requirements should be made that would allow schools to get credit for gains in achievement as well as absolute performance in a given year. The recently introduced pilot program that allows the use of longitudinal growth models by a small number of states is a step in that direction, but the fact that the unrealistic 100% proficiency requirement in 2013-2014 is maintained undercuts the value of the program. It is limited to just 2 states in 2006-2007 with the possibility of adding 6 more in 2007-2008. Furthermore, gains made by schools in states without longitudinal tracking systems do not count toward making AYP if the school fails to meet the AMO for either reading/English language arts or mathematics.

The narrow measures used to determine AYP should be expanded to encourage attention to subjects in addition to reading and mathematics and thereby avoid some of the narrowing of the curriculum that has been prevalent since NCLB was signed into law. Greater use should also be made of multiple measures within each of the subject areas assessed by augment state assessment results by taking into account information obtained from district benchmark assessments and classroom-based formative assessments.

Finally, the multiple-hurdle approach used to determine AYP should be replaced by a compensatory or hybrid approach. This would make the system fairer for schools that serve heterogeneous student bodies. It would also enhance the reliability of school classification.
References


Improving America’s Schools Act of 1994, Public Law No. 103-382.


Figure 1
Percent Proficient or Above on NAEP Grade 4 Reading in 2003 and 2005 for the 50 States
Figure 2
Percent Proficient or Above on NAEP Grad 8 Reading in 2003 and 2005 for the 50 states
Figure 3
Percent Proficient or Above on NAEP Grade 4 Mathematics in 2003 and 2005 for the 50 States
Figure 4
Percent Proficient or Above on NAEp Grade 8 Mathematics in 2003 and 2005 for the 50 States

Percent Proficient or Above on NAEp Grade 8 Mathematics in 2003 and 2005 for the 50 States

Year

2003

2005

Percent

0

10

20

30

40

50

60

70

80

90

100