"It's amazing to me how ridiculous this is. It's almost as if everybody has been set up to fail."

– Brenda Montoya, Las Vegas parent

"The ESEA [No Child Left Behind Act] is like a Russian novel. That's because it's long, it's complicated, and in the end, everybody gets killed."

— Scott Howard, former superintendent, Perry, Ohio, public schools

I. Set up to Fail

As states and districts tally test results and compile long lists of schools that have failed to make "adequate yearly progress" (AYP), the prospect of most U.S. public schools facing sanctions under the federal No Child Left Behind (NCLB) law seems increasingly inevitable. In just the first two years of this marathon race toward 100 percent proficiency, one quarter of all U.S. schools have already failed to make AYP (Center on Education Policy, 2004). Florida had the most schools on the failing list, with 88 percent (Miller, 2003). But Alaska, Delaware, Missouri, North Carolina and Oregon all have had 50 percent or more of their schools labeled as failing to meet targets for improvement (eSchool News, 2003).

In just the first two years of this marathon race toward 100 percent proficiency, one quarter of all U.S. schools have already failed to make AYP.

Whether or not NCLB's authors intended to set up the vast majority of public schools for failure, state takeover and possible private management, most observers agree that, barring substantive changes to the current law, this will be the likely result of the requirement that all students score "proficient" on state tests by 2014.

Not surprisingly, a large and growing number of those whose schools have landed on these AYP warning lists are criticizing NCLB for a range of defects. Families whose children attend public schools are receiving confusing and contradictory messages about their schools, rather than the clear and useful information promised by President Bush. In many cases, Florida being a prime example, schools showed consistent and marked improvement in state rankings, yet were judged to have failed when subjected to the NCLB formula.

Families whose children attend public schools are receiving confusing and contradictory messages about their schools, rather than the clear and useful information promised by President Bush.

The public at large (those with no direct involvement in public schools) is also seeing a somewhat confusing but predominantly negative portrait of public education. Daily newspapers report long lists of

"failing" public schools. They also feature school officials disputing the results, expressing exasperation, or pleading for a more balanced and nuanced view of how the schools are doing—not to mention more money to run their schools and administer the tests in a time of fiscal austerity.

Harvard Graduate School of Education Professor Richard Elmore is among those who say that NCLB's AYP provision is ungrounded in any proven theory of how schools actually improve. "The process of genuine improvement does not occur in equal annual increments. The AYP requirement, a completely arbitrary mathematical function grounded in no defensible knowledge or theory of school improvement, could, and probably will, result in penalizing and closing schools that are actually experts in school improvement" (Elmore, 2003).

"The AYP requirement, a completely arbitrary mathematical function grounded in no defensible knowledge or theory of school improvement, could, and probably will, result in penalizing and closing schools that are actually experts in school improvement"

-Richard Elmore

Rather than provide substantive answers to the many questions raised about NCLB, the U.S. Department of Education (DOE) launched a major public relations campaign to counter NCLB "negativism." The DOE allocated $500,000 to assemble a team to run the No Child Left Behind "Communications and Outreach" operation. The group, headed by a former Bush campaign operative, was tasked with supplementing existing DOE communications work. As criticism of the law reached a crescendo, Education Secretary Rod Paige and President Bush himself went on a public relations blitz to repair NCLB's flagging reputation. But no PR campaign can compensate for the deep flaws in the law, nor for the lack of adequate resources for struggling schools.

In a democratic system that depends on the contributions of all its citizens to the funding of public schools, NCLB's inaccurate picture of widespread, consistent failure is itself a major threat to the future of public education. The underlying structural dynamic of NCLB produces a vastly distorted picture, tarring both successful and underperforming schools as failures. Because of this distortion, NCLB risks not only undermining crucial support for public schools, but making it impossible to determine which schools and individual students really need substantial support and/or guidance on how to foster academic success.

NCLB risks not only undermining crucial support for public schools, but making it impossible to determine which schools and individual students really need substantial support and/or guidance on how to foster academic success.

This chapter will look at the predictions of massive failure to meet NCLB's targets for school improvement, show how those predictions

have been borne out, explain how such failure was built into the design of NCLB, and then illustrate what's wrong with relying on test scores alone to assess and improve public education.

A. The AYP Mess

The Adequate Yearly Progress (AYP) provision has many flaws:

- Widespread failure of schools to meet AYP targets was predictable and was indeed predicted by many experts who analyzed the law's provisions.
- After just two years, predictions of failure have been borne out in long lists of schools and districts landing on watch lists and lists of schools "in need of improvement."
- There are so many ways to fail under NCLB that it is difficult to draw comparisons between one failing school and another.
- High-poverty schools and districts are overwhelmingly the first to be identified as failing to make AYP.
- Diversity itself is penalized by the AYP formula. The more subgroups a school has, the less likely it will be to make AYP.
- Even well-off suburbs are not immune from failure if their schools include groups of students that struggle to perform well on state tests.

Widespread Failure: Predicted and Predictable

Once analysts began to digest NCLB's intricate provisions, it became clear even before the bill passed that high rates of school failure were the logical outcome of the law's approach to assessment and accountability. David Shreve, of the National Conference of State Legislatures, reflected the consensus of researchers when he estimated that 70 percent of all schools would be labeled "in need of improvement" in the coming years (Prah, 2002). State projections varied based on the difficulty of state tests, the rate of improvement expected, and the size of the subgroup chosen by the state as "statistically significant," but many projected massive levels of failure.

California's prospects for failure were perhaps most extreme. Under its performance standards, 98 percent of all schools in the state and 100 percent of schools serving mostly low-income students were expected to fail to meet the AYP goal. State Education Secretary Kerry Mazzoni explained, "We would rather set the bar high and not have everyone reach it than set it low and have everyone

David Shreve, of the National Conference of State Legislatures, reflected the consensus of researchers when he estimated that 70 percent of all schools would be labeled "in need of improvement" in the coming years.

State Education Secretary Kerry Mazzoni explained, "We would rather set the bar high and not have everyone reach it than set it low and have everyone reach it"

reach it" (Helfand, 2003). California's AYP plan required 7 percent per year gains, but in 2002 the state's actual test score gain was only about 1.5 percent. In a July 2003 report, the state Department of Education said just 32 percent of California schools achieved adequate progress for the year, lending support to dire predictions. Other states expecting over 90 percent of their schools to "fail" include Maine and Massachusetts, with Louisiana projecting 85 percent (Maine Education Association, 2003).

In 2002, a group of researchers met to discuss the AYP formula and to predict outcomes. Edward Haertel, of Stanford University, noted that if progress were based on the experience of the National Assessment of Educational Progress (NAEP) test results over the years, it would take 110 years to reach 100 percent proficiency across the country (Linn *et al.*, 2002c).

Edward Haertel, of Stanford University, noted that if progress were based on the experience of the National Assessment of Educational Progress (NAEP) test results over the years, it would take 110 years to reach 100 percent proficiency

Some proponents of NCLB argue that the law's "safe harbor" provisions will give schools breathing room. "Safe harbor" applies to subgroups that do not make AYP, if the percentage of students in that group decreases by ten percent from the previous year and that group made progress on another academic indicator. Others have pointed to the use of "rolling averages" (i.e., averaging scores over several years) as a means of reducing the impact of not doing well in one year. However, an analysis of state scores in Maine and Kentucky from the 1990s found that rolling averages will have very little impact, and safe harbor only a modest impact. Researcher Jaekyung Lee (2004) concluded, "Contrary to some expectations, the applications of both options would do little to reduce the risk of massive school failure due to unreasonably high AYP targets for all student groups."

If anything, predictions underestimated the extent of the failures and the confusion caused by constantly changing lists of "failing" schools.

INOI Lists Confirm Dire Predictions

Any hopes that predictions of large-scale failures were inaccurate or exaggerated were dashed by state-by-state lists of schools dubbed "in need of improvement" (INOI) based on state test results from the 2002-2003 school year (see Table I-1 at end of chapter). If anything, predictions underestimated the extent of the failures and the confusion caused by constantly changing lists of "failing" schools. A national teachers union estimated that 26,000 of the nation's 93,000 public schools failed to make adequate yearly progress in 2004.

- In New York City, 40 percent of the schools were labeled failing (Gootman, 2003).

- In New Mexico, more than 70 percent of the schools statewide would have failed if the new standards had been applied, so it got an extension of its compliance deadline from the U.S. DOE (Hutton, 2003).
- Fifty-seven percent of Delaware's public schools failed to make adequate yearly progress in math and reading, with 25 of the state's 28 high schools rated as under academic review, including one school that *Newsweek* magazine recently touted as among the best in the nation. At the middle school level, only three of the state's 33 middle schools made adequate yearly progress (Fuetsch, August 12).
- Just 32 percent of California schools achieved adequate progress. Elementary schools fared the best, with 37.2 percent meeting the benchmarks, but the achievement rate dropped to 19.7 percent in middle and high schools.
- In Mississippi, Pascagoula Schools Superintendent Hank Bounds said the new AYP guidelines may mean every school in the district might be classified as failing by the federal government — even the highest performing ones.
- In Michigan, 896 of the state's 3,472 public schools did not make AYP (Jacques, 2004).
- Lists of failing schools turned out to be fluid in Illinois, Texas and Minnesota. After an appeals process, Minnesota moved nearly half the schools, 93, off the failing list (School Funding Services, 2004).

In Mississippi, Pascagoula Schools Superintendent Hank Bounds said the new AYP guidelines may mean every school in the district might be classified as failing by the federal government — even the highest performing ones.

"Failure lists" are in flux for a number of reasons. The unprecedented flood of data is bound to include human and other errors, so that some schools are erroneously included or excluded. There is also the temptation on the part of schools, districts, and even state education agencies to underreport numbers of failing schools, or at least spin the numbers in the most positive way.

"Failure lists" are in flux for a number of reasons. The unprecedented flood of data is bound to include human and other errors, so that some schools are erroneously included or excluded.

A Massachusetts Department of Education press release, for example, emphasized statewide results showing that 94 percent of students made AYP (Massachusetts DOE, 2003). This looked much more positive than the fact that 67 percent of districts in the state were failing to make AYP because of the performance of one or more subgroups and were therefore on the path to possible NCLB sanctions (for a list of sanctions, see "Introduction: No Child Left Behind Testing and Sanctions Provisions"). The DOE press release did include the 67 percent statistic, but buried it near the bottom of the page.

Failure Comes in All Shapes and Sizes

Beyond sharing the stigma of failure and the risk of sanctions, many schools failing to meet AYP targets have very little in common with one another. Some schools clearly are failing to provide what their students need to be successful in higher education, life and work, although for a range of different reasons. Many are making improvements and progress, but not at a rate considered fast enough. Some have limited resources but nevertheless offer good educations to students who come to schools with enormous and growing needs. The vast majority of the schools with the farthest to go are in high-poverty urban communities**.** But NCLB also fails suburban schools rich in resources that comply with all but one or two of the law's many mandates.

Beyond sharing the stigma of failure and the risk of sanctions, many schools failing to meet AYP targets have very little in common with one another.

High-Poverty Schools Fail First

To virtually no one's surprise, high-poverty urban school districts are characterized by high, in some cases nearly universal, failure to make AYP. A report released in July 2003 by Michigan State University's Education Policy Center, for example, found**,** "Nearly all of Michigan's most troubled schools are in high-poverty urban areas and serve low-income, minority children." The report found that only seven out of the 216 troubled schools were in suburban and rural areas. David Plank, director of the center, said it is no secret that the poorest city schools perform most poorly, but the proportion—97 percent in this case — was more dramatic than expected (Putnam, 2003).

In Connecticut's poorest cities, there was widespread failure; for example, all of Hartford's high schools failed to make AYP

Replicating this study in other states would likely produce similar results. In Rhode Island, for example, the vast majority of schools needing improvement are concentrated in the urban districts of Central Falls, Newport, Pawtucket, Providence and Woonsocket. According to the *Providence Journal Bulletin*, "every urban middle school, which serves sixth through eighth graders, is in need of improvement" (Borg, 2003). In Connecticut's poorest cities, there was widespread failure; for example, all of Hartford's high schools failed to make AYP (Frahm, 2003).

Proponents of the NCLB approach to reform argue that it was precisely the intent of the law to highlight the failure of schools to serve low-income minority students. They say it is good that NCLB is shining a light on these failures because they can now be addressed.

But what does NCLB do to address these persistent problems other than punish kids, teachers, schools and communities?

Punishing Diversity

School diversity in and of itself can be another liability under NCLB (Doyle, 2003). Economists Thomas J. Kane and Douglas O. Staiger (2001) have found that racially integrated districts will be most likely to be found wanting because of significant racial disparities in test scores. Ironically, Kane and Staiger predict, districts that have gone out of their way to integrate are likely to be sanctioned more frequently than segregated school districts. In their study, Kane and Staiger looked at states that use racial subgroup test performance to determine ratings and found that segregated schools were less likely to suffer the consequences of score variability. This is largely because the number of students in any racial group within an integrated school is likely to be so small as to make scores for the subgroup more volatile than scores for the school as a whole.

Ironically, Kane and Staiger predict, districts that have gone out of their way to integrate are likely to be sanctioned more frequently than segregated school districts.

In California, for example, Kane and Staiger found that more diverse schools were substantially less likely to be rewarded for their test score gains than were more homogeneous schools, even though the more diverse schools actually had "greater improvements in overall test scores." Thus, use of test score gains to reward or punish "can generate perverse incentives for districts to segregate their students."

As Kane and Staiger predicted, schools *are* being punished under NCLB for being racially integrated. A report from Policy Analysis for California Education (Novak & Fuller, 2003) shows clearly that the more subgroups a school has, and the more economically disadvantaged students it enrolls, the less likely it is to make AYP.

Schools with very similar average scores fare very differently depending upon the number of subgroups they must report.

The report shows that schools with very similar average scores fare very differently depending upon the number of subgroups they must report. In California elementary schools with 50 to 75 percent economically disadvantaged students, 71 percent of those with three subgroups made AYP, while only 55 percent of those with five subgroups did. Yet the schools with more poor kids averaged one point higher in reading and had the same average math scores. In the elementary schools serving the most low-income students, the chance of making AYP fell 30 percentage points from schools having two subgroups to those

having five subgroups (from 64 to 34 percent). Yet, on average, the latter group scored only two points lower on the state reading test and one point higher on the math exam. In short, schools performing as well as other schools are failing simply because they are more diverse.

Defenders of NCLB argue that often the reason for the difference is that while a school may be serving some students well, it may serve others less well. While this has been shown on occasion, what Kane and Staiger and the PACE report demonstrate is that much of the failure to make AYP is an artifact of NCLB's formula. For example, a low-income, limited English proficient Latino child with special needs who scores low because the child has limited English proficiency will be counted in four subgroups. Just a few such children can cause an entire school to "fail." A higher-scoring, white, English-speaking, non-poor student counts just once.

A low-income, limited English proficient Latino child with special needs who scores low because the child has limited English proficiency will be counted in four subgroups.

Measurement expert Robert Linn confirmed Kane and Staiger's findings that the requirement for sub-group AYP will make more schools vulnerable to being labeled "failures." Linn concludes, "The NCLB adequate yearly progress requirements represent enormous, if not overwhelming, challenges to schools, districts, and states" (Linn, 2003b).

"The NCLB adequate yearly progress requirements represent enormous, if not overwhelming, challenges to schools, districts, and states"
-Robert Linn

Reports from the field demonstrate that school officials in urban districts are feeling the burden of this "diversity penalty." A study by the nonpartisan Center on Education Policy highlighted the challenge faced by urban districts with diverse, low-income student populations in a report titled "*Implementing The No Child Left Behind Act*" (Center on Education Policy, 2003). While the report in general asserted that districts are "optimistic" about their ability to meet the requirements of the law, urban districts in particular see serious obstacles to their success. "These urban systems faced special challenges in making adequate yearly progress because they tended to have more student subgroups counted for accountability purposes and more schools targeted for improvement and technical assistance. For 2003-04, Cleveland has 21 schools identified for school improvement or corrective action. To make AYP, the Cleveland public schools must show improvement every year on all 82 benchmarks in the state's AYP definition—taking into account all the subgroups, grade levels, and progress indicators counted—far more than its suburban counterparts. And because this AYP definition is based in part on state average test scores, dis-

tricts with low performance, like Cleveland, must make up more ground than other districts in the state to meet the yearly benchmarks."

Massachusetts superintendents interviewed for a report by MassINC (2003), a business-backed research group that strongly supports test-based accountability and reform, also highlighted the problem of the diversity penalty. The report's authors write: "Superintendents noted their conviction that, because of their disproportionate impact, student subgroups with lower achievement rates are at risk of alienation, blame, and damaged self-confidence."

"Because of their disproportionate impact, student subgroups with lower achievement rates are at risk of alienation, blame, and damaged self-confidence."
- Massachusetts Superintendents

Failure in the Suburbs

While few observers express surprise at seeing large numbers of high-poverty urban schools show up on lists of failing schools, the struggles of suburban schools and districts to keep up with AYP's demands is another matter. Some NCLB proponents see suburban school failures as more proof that the law is living up to its name by identifying the pockets of poorly performing students who might otherwise go unnoticed when their scores are averaged with those of their high-performing peers. Critics question the accuracy and efficacy of tarring entire schools and districts because of the performance of specific subgroups.

Suburban school officials complain that, rather than provide support, guidance and resources to address the needs of vulnerable students in these subgroups, NCLB lays out traps and pitfalls for both excellent and neglectful schools. Some suburban districts fall victim to the diversity penalty. Many suburban schools are tripped up by the requirement that 95 percent of all students (and 95 percent of students in each subgroup) be tested. This stipulation, in particular, has led to schools that would otherwise meet and exceed expectations for improvement being labeled INOI because a handful of students were not tested. One Georgia school, for example, was labeled INOI because a single disabled student missed the state math test, meaning the school had only 42 out of its 45 special ed students take the test for a participation rate of 93, not 95 percent. If that student had taken the math exam, the participation rate would have been 96 percent and the school would have made its goals (Tofig, 2003).

Rather than provide support, guidance and resources to address the needs of vulnerable students in these subgroups, NCLB lays out traps and pitfalls for both excellent and neglectful schools.

A *Contra Costa Times* computer analysis of California schools failing to make AYP found that one-third of schools and districts in Contra Costa, Alameda and Solano counties failed to make AYP because they failed to meet the 95 percent participation rate, more than for any academic reason. In the Pleasanton schools, for example, if three more English learners and two more Latino students had taken the tests, the schools would have succeeded in staying off the list of schools that had failed to make AYP (Pardington, 2003).

In Brookline, Massachusetts, an affluent yet diverse town bordering Boston, if one more Hispanic student had taken the state test, the town's one high school would have made AYP (Holland, 2003).

In Brookline, Massachusetts, if one more Hispanic student had taken the state test, the town's one high school would have made AYP

Conclusion

Well before its passage, those who analyzed NCLB anticipated the AYP train wreck. Now public school families are paying the price in dislocation and confusion. When schools are labeled inadequate based on the statistical idiosyncrasies of the AYP formula, the result can be the opposite of providing meaningful accountability to parents and the community. As one parent put it, "Seeing the voluminous information generated by the 'No Child Left Behind' scores left me with an unaccustomed feeling: that I had way too much information about Kentucky's schools. Unfortunately, the wealth of information provided little insight about what I, as a parent and taxpayer, am supposed to do to make schools better" (Truman, 2003).

"Seeing the voluminous information generated by the 'No Child Left Behind' scores left me with an unaccustomed feeling: that I had way too much information about Kentucky's schools. Unfortunately, the wealth of information provided little insight about what I, as a parent and taxpayer, am supposed to do to make schools better"

-Kentucky Parent

There is a widespread consensus among researchers, educators, parents and others about the mechanics of AYP:

- Widespread failure was an inevitable outcome of its design and is being borne out in school and district results.
- There are many ways to fail under AYP, so many different kinds of schools are labeled failures.
- High-poverty schools and districts fail first, but diversity itself is punished. Failure afflicts well-off suburbs as well.
- School officials are well aware of the various ways AYP trips them up and feel frustrated rather than empowered to initiate or continue efforts toward positive change.

B. What is "Proficiency"?

While the common perception holds that universal proficiency is a reasonable and desirable goal, it is important to understand the basis for the use of the term "proficient" in NCLB and how this could contribute to massive school failure. NCLB aims for 100 percent student proficiency by 2014. Who could object to the demand that all public schoolchildren be "proficient" in math and English? The answer depends on how you define the word. The term "proficiency" is borrowed from the National Assessment of Education Progress (NAEP) tests, which have been used for years to provide sampled snapshots of how U.S. schoolchildren are progressing in various academic subjects. NAEP's importance will grow enormously as a result of NCLB, which requires states to use NAEP to confirm the results of their state tests.

Many newspaper reports and commentaries on NCLB and NAEP describe proficiency simply as the ability to do grade-level work. For example, a recent *Boston Globe* article on NAEP results included this explanation of what it means to score below the NAEP proficient level: "That means [students] struggle with grade-level reading and math and cannot always apply the skills to real-world situations" (Vaishnav, 2003).

A wide range of testing experts, however, have found that the definition of proficiency used by states for NCLB purposes is wildly inconsistent, and the NAEP proficiency standards are set so high that it will be impossible for most schools to reach them (Stecher *et al.*, November 2003; Linn, 2002b; Bracey, 2003). This section will address the following issues:

- The term proficiency comes from its use in NAEP testing, where it has been widely criticized for being an unrealistic and inaccurate standard.
- States vary wildly in how they define proficiency, making it difficult if not impossible to make meaningful comparisons from state to state.
- As a result, states are beginning the race to 100 percent proficiency from many different starting points, many of which do not necessarily reflect the relative academic health of their schools and students.
- Some states have resorted to lowering their standards in hopes of making the grade.

A wide range of testing experts, however, have found that the definition of proficiency used by states for NCLB purposes is wildly inconsistent, and the proficiency standards are set so high that it will be impossible for most schools to reach them

"Proficient" does not mean "grade level" or "average," contrary to common public interpretation.

NAEP Levels

Research on the setting of NAEP "levels" (basic, proficient, advanced) shows the levels to be misleading. For example, "proficient" does not mean "grade level" or "average," contrary to common public interpretation.

According to Professor Gerald Bracey, an independent education researcher who teaches at George Mason University, "[T]he NAEP achievement levels have been rejected by everyone who has ever studied them: UCLA's Center for Research on Evaluation, Student Standards and Testing (CRESST), the General Accounting Office and the National Academy of Sciences, as well as by individual psychometricians such as Lyle Jones of the University of North Carolina. The studies agree that the methods used are flawed and, more importantly, the results don't accord with any other data.

"For instance, Jones pointed out that American fourth-graders were well above average on the mathematics tests of the Third International Mathematics and Science Study (TIMSS), yet only 18 percent reached the proficient level and a meager two percent scored at the advanced level in the 1996 NAEP mathematics. Similar low percentages are seen in the 1996 NAEP Science assessment and TIMSS Science where American fourth-graders were third in the world among 26 nations. Finally, on the 2000 NAEP reading assessment, only 32 percent of fourth-graders attained proficient or better, but that [sic] American 9-year-olds were second in the world among 27 countries in the international reading study, *How in the World Do Students Read?* It makes no sense that American kids do so poorly on domestic measures such as NAEP but stack up well against the rest of the industrialized world" (Bracey, 2003).

A 1998 report from the National Academy of Sciences titled, "Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress," recommended that the process for setting NAEP achievement levels be replaced. "This committee, as well as the U.S. General Accounting Office, the National Academy of Education, and other evaluators, have judged the current achievement-level-setting model and results to be flawed...NAEP achievement-level results do not appear to be reasonable compared with other external information about students' achievement" (Pellegrino *et al.*, 1998).

"American 9-year-olds were second in the world among 27 countries in the international reading study, "How in the World Do Students Read? " It makes no sense that American kids do so poorly on domestic measures such as NAEP but stack up well against the rest of the industrialized world"

- Gerald Bracey

NAEP achievement-level results do not appear to be reasonable compared with other external information about students' achievement"

-National Academy of Sciences

Robert Linn and his colleagues (2002a) report that few states will meet AYP targets, precisely because the "proficient" level in most states is set too high. Perhaps more important is their contention that NCLB's demand for steady year-to-year progress toward 100 percent proficiency is completely unrealistic. They found, for example, that only three out of 33 states posted even a one percent per year increase in the number of students scoring proficient on National Assessment of Educational Progress reading tests from 1992 to 1998. This is far below the 5 or 6 percent increase per year that would be required to reach 100 percent proficiency by 2014.

There will be great variation in the percentages of students deemed proficient from state to state, even if the students have the same skills.

Proficiency Levels Vary Widely

State proficiency standards are not only unreasonably high, they are anything but "standard." The Northwest Evaluation Association (NWEA) did a careful analysis of proficiency standards in 14 states and documented wild variations both among states and within states from grade level to grade level (for example, math proficiency might be set at the 46th percentile in grade 3 and at the 75th percentile in grade 8). The study generated three conclusions: there will be great variation in the percentages of students deemed proficient from state to state, even if the students have the same skills; differences in proficiency standards across grades will provide teachers with inconsistent information about students; and variations in proficiency standards between subject areas will provide inconsistent information when comparing proficiency in math and English (Kingsbury *et al.*, 2003). Variations in proficiency levels from grade to grade could have dire consequences. For example, if the level of math proficiency is set lower in grade 3 than in grade 8, a student might not be identified as needing math help until 8th grade, when it may be much harder to intervene successfully.

A look at each state's AYP starting points, or where states begin the race to 100 percent proficiency, gives yet another indication of the variability of state proficiency standards.

A Race with 50 Starting Points

A look at each state's AYP starting points, or where states begin the race to 100 percent proficiency, gives yet another indication of the variability of state proficiency standards. Starting points are determined by using a formula based on the percentage "proficient" on state tests in 2001-02 (see Table I-2 at end of chapter). There is a startling degree of variability. Some states are faced with the challenge of reaching 100 percent proficiency in English and Language Arts by 2014 starting from lows like 13.6 percent for California high schoolers and

23 percent for high school students in Arizona. At the other end of the spectrum, Colorado students begin the race at 80.3 percent proficiency.

Are these starting points an accurate reflection of the relative academic health of these states? Professor Linn takes issue with the utility of such comparisons. In his report, "Performance Standards: Utility for Different Uses of Assessments," Linn wrote: "State NAEP results indicate that states do vary in terms of student achievement, but not nearly enough to explain the huge variability in NCLB percentage proficient starting points. For the 43 states that participated in the 2002 NAEP 4th grade reading assessment, for example, the percentage of students who were at the proficient level or above ranged from a low of 14 percent in Mississippi to a high of 47 percent in Massachusetts" (Linn, 2003a).

Under NCLB more than twice as many students in Wyoming could be identified as below proficient than in Montana, even if students in both states have exactly the same achievement level on a norm-referenced test.

The AYP gain requirements differ so markedly in part due to different state definitions of "proficient." The NWEA study (2003) compared state-set proficiency levels in the similar, adjacent states of Wyoming and Montana. The eighth-grade math proficiency level in Montana was set at the 36th percentile, while Wyoming's was set at the 89th percentile. This means that under NCLB more than twice as many students in Wyoming could be identified as below proficient than in Montana, even if students in both states have exactly the same achievement level on a norm-referenced test.

Shifting Standards to Make the Grade

Some states, including Texas, Arizona, Colorado and Louisiana, have lowered their definition of proficiency in hopes of increasing their chances of attaining AYP. Colorado and Louisiana created a dual system, with a lesser definition of "proficient" for meeting federal requirements and a more stringent definition for local use (FairTest *Examiner*, Fall 2002). Rod Paige, U.S. Secretary of Education, denounced these moves in a letter to state education superintendents, yet Colorado was one of the first states to have its NCLB plans approved by his agency, the U.S. Department of Education.

Some states, including Texas, Arizona, Colorado and Louisiana, have lowered their definition of proficiency in hopes of increasing their chances of attaining AYP.

In other cases, state accountability requirements are easier to meet than AYP requirements. For example, in Arizona, 289 schools were identified as needing improvement under NCLB, but these same schools met the state's performance targets and earned either a "performing" or "highly performing" label. In Virginia, 723 (40 percent of

all schools) failed to make federal AYP goals while only 402 (22 percent) failed to meet state accreditation standards (Sunderman and Kim, 2004). (See Appendix 1: "Jeb's A+ Schools Dubbed Failing by George," and Appendix 2: "When AYP Means Good Schools Are Called Bad" at the end of this chapter.)

Such dual standards are likely to increase confusion among parents about the meaning of the various ratings.

Such dual standards are likely to increase confusion among parents about the meaning of the various ratings. Dual systems could also increase administration costs if larger staffs will be needed to compile and analyze two sets of statistics.

Some state education policymakers could be holding onto the original state system because they are betting NCLB will be overhauled before 2014. "If you chuck your entire state system and later on the federal law does change, then you're left with an unworkable system," said William Padia, director of policy and evaluation at the California Department of Education. "Better to live with an uncomfortable marriage of the two" (Galehouse, 2003).

In response to the problems implementing NCLB during its first two years and after seeing that different states have widely varying compliance agreements with the federal government, more than 40 states have requested permission from the U.S. DOE to make changes in their NCLB accountability plans (Olson, 2004). Connecticut wants to test only in grades 4, 6, 8, and 10, rather than 3-8 inclusive. Other states plan to revise the minimum size of the "cells" used to determine whether a subgroup will be counted in AYP determinations. A number of states want to use "growth" models (student progress) rather than the absolute score requirements set through NCLB. Such approaches could allow schools to gain credit for students' partial progress toward the proficient level, slowing down the speed at which they approach "in need of improvement" status. Some states want to be able to count students who fit multiple subpopulations (e.g., Latino, limited English proficient, low-income and disabled) in just one subgroup for AYP calculations.

Some state education policymakers could be holding onto the original state system because they are betting NCLB will be overhauled before 2014.

Problems with the definition of "proficiency" mean that NCLB's goal of bringing all students up to "proficiency" is anything but reasonable.

- By borrowing from NAEP terminology and using the NAEP levels to double-check state progress, the nation begins with a false assumption about how well the majority of U.S. students are doing.
- The variability in defining proficiency means that schools, educators, students and their families face severe consequences in some states that they would not face in other states with similar school quality and student attainment.
- Some states are responding to the unreasonable demands of NCLB by lowering their standards in hopes of avoiding sanctions.

Errors are common in standardized testing and are likely to skyrocket with the explosion of testing brought about by NCLB.

C. What's Wrong with Reliance on Test Scores?

A key reason NCLB sets schools up to fail is its almost total reliance on standardized test scores to judge the success or failure of public schools. What's wrong with using test scores alone? A number of independent researchers have investigated this question and reached the following conclusions:

- Test score fluctuations are often a matter of luck rather than real progress.
- A certain level of failure is built into the design of most standardized tests.
- Errors are common in standardized testing and are likely to skyrocket with the explosion of testing brought about by NCLB.

Year-to-year gains and losses on state tests are too unreliable to be used for decision-making.

(The next chapter will analyze the dangers to curriculum and instruction caused by overreliance on standardized test scores.)

Score Swings a Matter of Luck

School officials and politicians trumpet each bump upward in test scores and respond to such movement with rewards and still more test-driven education policy. Reaction tends to be more muted when the scores that have gone up eventually come down again. Researchers Kane and Staiger (2001a, 2001b) warned about these problems even before NCLB passed Congress. Their research into the reasons for test score fluctuations, for example, consistently found that year-to-year gains and losses on state tests are too unreliable to be used for decision-making.

In two papers Kane and Staiger examined data, primarily from North Carolina, to determine the precision of school scores. They categorized test score variations as due to sampling changes (e.g., a different group of students each year in a tested grade), a particularly severe problem in small schools; "one-time factors" such as a barking dog that distracts a group of test-takers; and persistent differences in actual performance among schools.

The researchers found that 50 to 80 percent of the year-to-year observed fluctuation in a typical North Carolina school's average score is due to the first two factors, not differences in tested achievement.

School rankings based largely on score increases "generally resemble a lottery."
- Kane and Staiger

As a result, school rankings based largely on score increases "generally resemble a lottery." Only one percent of the state's schools ranked in the top 10 percent in math for all six of the years studied. In reading, which is more volatile than math, more than one-third of all schools ranked in the top 10 percent at some point.

Selecting "good" programs that other schools should emulate, a common goal of test-based accountability programs, is also a matter of luck. If test scores are the determining factor, a large percentage of schools will at some point be labeled "best practice" schools. The result would be an ever-expanding menu of "best practices" from schools whose scores often decline over the next year or two.

Selecting "good" programs that other schools should emulate, a common goal of test-based accountability programs, is also a matter of luck.

Another researcher, Boston College education professor Walter Haney, looked at all Massachusetts elementary schools that showed a 10-point gain on the state test from 1999-2000 and found that most posted score declines in 2001, often as large as the gains of the previous year. The data showed that a school that did better the first time was more likely than not to do worse the second time.

According to Haney, "These results don't mean that teachers or students became lazy and tried to coast on their success. They mean that there was never really evidence of success at all. Particularly in small schools, as other research has confirmed, changes in score averages from year to year are poor measures of school quality. If fewer than 100 students are tested in each grade, averages may swing widely from year to year simply because of the particular samples of students tested and the vagaries of annual test content and administration" (Haney, 2002).

If too few students are in a school or subgroup that must be measured for its progress, the results will be particularly volatile. However, most states use minimum subgroup sizes that are far smaller than those that Haney and other testing experts say must be used to attain statistically sound results. Among the states whose AYP plans were initially approved by the U.S. Department of Education, Colorado, Indiana and Ohio will use 30 as a minimum group size. Massachusetts will use only 20. (See Table I-3 and Chart I-1 at end of chapter.)

If subgroup sizes are too small, some schools will fall short of AYP due to measurement error, not because of any academic problem.

If subgroup sizes are too small, some schools will fall short of AYP due to measurement error, not because of any academic problem. Some states will use statistical procedures to ameliorate the effects of too-small groups. These states determine the margin of error for each subgroup, based on the number of students tested and the percent who reach proficiency, and use it to give themselves more leeway in determining how many schools have failed to make AYP. Kentucky is one state using a so-called confidence-interval formula. A *Lexington Herald-Leader* reporter described how such a process works: "If a Kentucky school had 17 African-American students and three of them — 18 percent — reached proficiency, the margin of error for that subgroup would be plus or minus 28 percentage points. Those points would be added to the raw score, and a school would be evaluated as if 46 percent of students had scored proficiently" (Deffendall, 2003). While this makes statistical sense and helps prevent over-identification of schools as failing, it typically mystifies the public, which often concludes states are manipulating numbers to hide failure.

Schools began tailoring their curricula to improve performance on the tests, without generating similar improvements on other measures.

In the hopes of making more accurate judgments, some states have taken to considering multi-year, moving averages rather than just the previous year's scores. However, as Kane and Staiger report (2001a, 2001b), a North Carolina school which desired to predict its current year reading score gains would be better off to simply pick the state's average score increase rather than to use its own previous four years of score changes. Thus, averaging a few years' scores in an effort to solve the problem of random fluctuations appears not to sufficiently reduce misinformation.

The authors also asked whether score gains are due to such factors as teaching to the test rather than to real improvements in learning. Lacking a direct measure, they examined several characteristics of "student engagement" -- absenteeism, time doing homework, and time watching television. Those measures did not improve in schools in

which scores rose substantially. The authors note that this lack of improvement "would be consistent with the hypothesis that schools began tailoring their curricula to improve performance on the tests, without generating similar improvements on other measures."

Tests like the MCAS [the Massachusetts Comprehensive Assessment System exams] are designed so that all students can never succeed.
-Walt Haney

Failure Built In to Test Design

Some state tests are constructed to guarantee that a certain percentage of students will fail every year. This is because the test design methodology commonly employed in state exams rests on a technology used to develop norm-referenced tests, which sort and rank test-takers, always leaving some students at the bottom of the curve. Low-scoring students are disproportionately low-income, African-American, Latino, recent immigrants whose first language is not English, and students with disabilities.

In "Ensuring Failure," Haney (2002) wrote: "When questions answered correctly by more than 70 percent of students are systematically excluded from the exam, this guarantees continuing failure. Tests like the MCAS [the Massachusetts Comprehensive Assessment System exams] are designed so that all students can never succeed. Evidence suggests that other state tests (in Texas, California, and New York, for example) also have been constructed using norm-referenced test-construction procedures." In fact, research conducted for a lawsuit challenging the Texas graduation test found the TAAS exam had the flaws described by Haney (FairTest *Examiner*, 1999-2000).

The sharp increase in testing volume created by NCLB may cause a spike in human errors unless greater attention is paid to quality control issues.

Error-Prone Tests

The New York Times (Henriques, 2003) reported in detail on the prevalence of errors in standardized testing. A series of articles warned that the sharp increase in testing volume created by NCLB may cause a spike in human errors unless greater attention is paid to quality control issues.

Prof. Mark L. Davison, an educational psychologist at the University of Minnesota, predicted a doubling of testing in the next few years as a result of NLCB. "I think preventing [errors] entirely is impossible," he told the *Times.* "As existing companies expand and new companies move into the field, they're going to experience growing pains."

A report by the National Board on Educational Testing and Policy based at Boston College, *Errors in Standardized Tests: A Systemic Problem*, highlights the nature and extent of human mistakes in educational testing over the past 25 years (Rhoades and Madaus, 2003). In contrast to random measurement error expected in all tests, human error is unexpected and brings unknown, often harmful consequences for students and schools, including:

- Inaccurately preventing high school seniors from receiving a diploma (Minnesota 2000);
- Creating misleading "worst school" lists (Pennsylvania 1996, Nevada 1999, Ohio 2002);
- Erroneously assigning students to remedial classes or retaining them in grade (New Jersey 1993, New York City 1999, Maryland, 2001);
- Barring qualified college applicants from attending their chosen universities (Scotland 2000, England 2002); and
- Denying qualified applicants access to professional credentials (Alabama 1981-85, New York 1981, Oklahoma 2000).

These errors occur in an industry whose activities are largely unregulated, an environment where mistakes are difficult to detect.

Rhoades and Madaus point out that these errors occur in an industry whose activities are largely unregulated, an environment where mistakes are difficult to detect. As the amount of testing has increased, the industry has been spread thin and testing errors have risen. NCLB's mandated increase in testing is likely to cause a larger jump in the number of errors in designing tests, setting passing scores, establishing norm groups, scoring exams, and reporting results.

The Boston College report demonstrates that testing is a fallible technology, subject to internal and external errors. With errors an unavoidable problem, basing important educational decisions on the outcome of one test can put children and schools at risk due to foul-ups that may never be caught or remedied.

The release of Illinois's NCLB report card — with an estimated 34,261 mistakes involving about 75 percent of the state's schools, according to the *Chicago Tribune* — offers a vivid illustration of the Boston College report's conclusions. The Tribune said there were so many errors, "it is virtually impossible to draw meaningful conclusions that educators had hoped for and that the federal No Child Left Behind Act requires." State school officials said 368 schools may have been mistakenly placed on a federal failure list because of data errors.

There were so many errors, "it is virtually impossible to draw meaningful conclusions that educators had hoped for and that the federal No Child Left Behind Act requires."

-Chicago Tribune

The list reportedly included several well-known, high-performing elementary and high schools (Banchero and Little, 2003)

Summary

For President Bush and other NCLB proponents, the law's near-total reliance on test scores to determine the progress of students, teachers and schools reflects a desire for "objective" assessments of how schools are doing. Bush says, for example, "Without yearly testing, we don't know who is falling behind and who needs help. Without yearly testing, too often we don't find failure until it is too late to fix" (Bush, 2001). But standardized test scores, while they have some utility, offer nothing more than a snapshot of student achievement at a moment in time and can be misleading when used to make important decisions about students and schools. The national focus, some would say obsession, on standardized test scores to drive school improvement and reform is not an entirely new policy. The historic record casts serious doubt on the decision to continue a nationwide experiment in test-based reform. Among the findings:

- Test score fluctuations do not necessarily indicate real progress when scores rise or deterioration when they fall and should not be used by themselves to reward or sanction schools, teachers or school officials.
- Many of the tests that are being used to judge our students, teachers and schools are specifically designed to ensure a certain proportion of failures.
- Errors have always been a part of standardized testing and are likely to increase substantially with the increase in testing mandated by NCLB.

D. Conclusion

Most NCLB critics do not deny that there are public schools offering inadequate educations to their students. On the contrary, they say, there are still too many schools that, for a variety of reasons, are not giving students what they need to succeed. NCLB, however, will exacerbate this situation by labeling so many schools as underperforming that it will be impossible to really identify schools that need improvement and give them what they need to improve. As the National Education Association's Joel Packer put it, "It's like saying everybody in the country is sick. How do you figure out how to focus

"It's like saying everybody in the country is sick. How do you figure out how to focus your resources, especially if what you want to do is help those who are really sick?"
-National Education Association

"The adequate yearly progress [AYP] net has been cast very wide, and so it's going to catch a very high percentage of schools"

your resources, especially if what you want to do is help those who are really sick?" (National Education Association, 2003).

David Shreve, an education policy expert at the National Conference of State Legislatures, concurs with this view (Marks, July 21). "The adequate yearly progress [AYP] net has been cast very wide, and so it's going to catch a very high percentage of schools," says Shreve. "We either have to accept the fact that a vast majority of our schools are awful, or we have to accept the AYP net has been cast too broadly and it's catching way more schools than it should."

"There's no guarantee whatsoever under No Child Left Behind that any school has the basic resources that they need to bring these children up to the level of achievement the law calls for,"
-Michael Rebell

While the AYP formula is mind-bogglingly complex and will have a range of complicated consequences for public schools, behind it is a simplistic view that most educators are not working hard and that improvement will result when they feel the "tough love" of a kick in the pants.

Most of those who work in and attend these schools see a vastly different reality: many challenges and dwindling resources. The people who are charged with reaching the 100 percent proficient goal see a law that is heavy on punishment and light on the tools they need to do the job. "There's no guarantee whatsoever under No Child Left Behind that any school has the basic resources that they need to bring these children up to the level of achievement the law calls for," says Michael Rebell, executive director and counsel of the Campaign for Fiscal Equity, a nonprofit education funding advocacy group in New York. "You can beat them over the head as much as you want, but you can't get blood from a stone" (Marks, 2003).

Or, as the principal of a predominantly black North Carolina charter school put it: "I'm not so sure this law is about leaving no child behind." Jackie M'Buru, principal of SPARC Academy in Raleigh, added, "I think this law is about blaming teachers and principals who need more support. If you say you need more help, the answer is: 'Get better or we'll shut you down'" (Simmons, 2003).

"I think this law is about blaming teachers and principals who need more support. If you say you need more help, the answer is: 'Get better or we'll shut you down'"
-Jackie M'Buru

Appendix 1: Jeb's A+ Schools Dubbed Failing by George

If anyone was going to get President George W. Bush's NCLB test right, you'd think it would be his brother, Florida Governor Jeb Bush. So are Florida's schools models of the improvements to be reaped by test-driven reform? That depends on which Bush brother you ask (Pinzur, 2003).

Florida Governor Jeb Bush touted improvement based on Florida's ranking system, with six times as many A schools in 2003 than when grading began in 1999 and fewer than half as many F schools. But within months, President George Bush's NCLB report card found that of 1,229 A schools statewide, only 22 percent made Adequate Yearly Progress.

Six percent of the state's B schools made AYP, as did two percent of C schools. No D or F schools qualified. Florida education officials scrambled to explain. 'Just like an `A' student has room for improvement, even a top school can work toward improving performance,'' said Frances Marine, spokeswoman for the Florida Department of Education. But the contradictory picture produced by the state and federal ranking systems was a perfect illustration of how unenlightening it is when accountability is boiled down to standardized test results. Parents are left struggling to make sense of it all.

Education leaders are especially concerned about confusing parents, who have heard the governor celebrate annual improvements in school grades. "If I saw that my state graded me as an `A' and then the federal government said we hadn't met the No Child Left Behind Act, I would be very confused and asking a lot of questions,'' said Karin Brown, a parent activist and former president of the Dade County Council PTA/PTSA. ``From a parent point of view, there's definitely a contradiction here.''

An editorial writer for the Bradenton Herald summed it up: "The answer, of course, is that neither represents a fair and accurate picture of the quality of education being provided to students here or elsewhere in the state, which had a 90 percent failure rate in the federal test. Rather, they are snapshots of student performance as measured by an arbitrary set of standards. But they don't necessarily reflect teachers' success at educating children with widely varied levels of ability, socialization and language mastery. Throw in a different set of standards and you likely would get yet another, entirely different result" (Editorial, 2003).

The contradictory picture produced by the state and federal ranking systems was a perfect illustration of how unenlightening it is when accountability is boiled down to standardized test results.

Appendix 2: When AYP Means Good Schools Are Called Bad

The rigidity, complexity and insatiable demands of NCLB's AYP formula mean that schools across the country that had been lauded for improvement and excellence have unceremoniously landed on lists of schools "In Need of Improvement."

One was Southfield, Michigan's Vandenberg Elementary, which President Bush had visited to promote ESEA. *USA Today* found that 19 U.S. Department of Education Blue Ribbon exemplary schools ended up on low-performing lists (Thomas and DeBarros, 2003).

USA Today *found that 19 U.S. Department of Education Blue Ribbon exemplary schools ended up on low-performing lists*

Mark Christie, the Republican former president of the Virginia Board of Education, has decried the way NCLB labels many excellent schools as INOI. "If you create a system that calls a good school a bad school, people will know and lose faith in accountability," Mr. Christie said.

Christie was quoted in a New York Times article by Michael Winerip (2003) that highlighted the experience of one Virginia school, Tuckahoe Middle School in suburban Henrico County. According to Winerip, "Tuckahoe's test scores are among the best in Virginia, with 99 percent achieving proficiency in math, 95 percent in English. Its previous principal was the 2002 state principal of the year, and in 1996 Tuckahoe was named a Blue Ribbon School of Excellence by the federal Education Department." How did such a school end up on the list of schools INOI? It missed by one percentage point the mandate that 95 percent of students be tested because it had recent Bosnian immigrants who didn't speak English well enough to be tested in English and the state did not have a test to give them as an alternative. "It didn't make sense to have them take a test they couldn't understand," said Kurt Hulett, Tuckahoe's principal.

Kentucky found more than a quarter of its schools, 470, failed to make AYP last year. Forty of those same schools had been recognized just one month earlier for being a year ahead of their academic goals.

Of course, Tuckahoe is far from an isolated example. In Tennessee, about 160 schools received incentive grants of $5,150 last year for meeting state benchmarks. A few months later, 40 of those schools wound up on this list of schools that failed to make AYP (Riley, 2003).

Kentucky found more than a quarter of its schools, 470, failed to make AYP last year. Forty of those same schools had been recognized

just one month earlier for being a year ahead of their academic goals set by the Kentucky state testing system. Emma B. Ward Elementary School was one of the schools recognized for improvement by the state and called failing by the feds. School officials complained that the federal designation was out of date and did not reflect the improvements they had already made. Principal Sarah Sweat said, "It's really unfortunate. We know that what we did last year got us to the score we got, which was the highest in the district. We know we are doing the right thing to achieve our goals" (Rodriguez, 2003).

The feds have come up with a solution to the embarrassing problem of Blue Ribbon Schools being labeled failures by NCLB. While in the past, Blue Ribbon Schools were evaluated on multiple measures, including school visits and interviews with support staff, now getting a Blue Ribbon will depend solely on test scores.

While in the past, Blue Ribbon Schools were evaluated on multiple measures, including school visits and interviews with support staff, now getting a Blue Ribbon will depend solely on test scores.

References

Banchero, S. and Little, D. December 19, 2003. "Errors Fill State Testing Data," *Chicago Tribune.*

Borg, L. October 10, 2003**. "**Schools Decline in Annual Rankings," *Providence Journal Bulletin*.

Bracey, G. February 2003. "The No Child Left Behind Act: Just Say No." Available online at http://www.nochildleft.com.

Bradenton (FL) Herald. August 10, 2003. "Accountability Sham: State, Federal Tests Results Conflict."

Bush, G. W. January 23, 2001. Press Conference with President Bush and Education Secretary Rod Paige to Introduce the President's Education Program. Available on the web at http://www.whitehouse.gov/news/releases/2001/01/20010123-2.html.

Center on Education Policy. October 2003**. "**Implementing The No Child Left Behind Act." Available online at www.cep-dc.org.

Center on Education Policy. January 2004. "From the Capital to the Classroom." Available online at www.cep-dc.org.

Deffendall, L. October 19, 2003. "No Statistics Are Being Left Behind," *Lexington Herald-Leader*.

Dillon, S. January 2, 2004. "Some School Districts Challenge Bush's Signature Education Law," *The New York Times.*

Doyle, D. P. April 16, 2003. The Doyle Report, Issue 3.16, No. 63. Available online at http://www.thedoylereport.com/cyber_chair?object=archive[]&content_id=3696.

Editorial. August 10, 2003. "Accountability Sham: State, Federal Test Results Conflict," *Bradenton Herald* (FL).

Elmore, R. F. November 2003. "A Plea for Strong Practice," *Education Leadership*, Volume 61, Number 3.

eSchool News. October 1, 2003. "A Look at School AYP Failure by State." Available online at http://www.eschoolnews.com/news/issue.cfm?PubID=1&IssueID=201.

FairTest *Examiner*. Winter 1999-2000. "Court Rules for High-Stakes Testing." Available online at http://www.fairtest.org.

FairTest *Examiner*. Summer 2002. "ESEA: Ten Percent of U.S. Schools Labeled 'Failing.'" Available online at http://www.fairtest.org.

Frahm, R. A. December 18, 2003. "46% of Schools `Left Behind'," *Hartford Courant.*

Fuetsch, M. August 12, 2003. "Delaware Schools Reeling**:** Unforgiving Federal Ratings Fail to Show Actual Progress," *The News Journal.*

Galehouse, M. July 14, 2003. "Academic Bar Lowered to Get Schools on Track," *The Arizona Republic.*

Gootman, E. September 11, 2003. "40 Percent of City Schools Do Not Meet U.S. Standards," *The New York Times.*

Haney, W. July 10, 2002. "Ensuring Failure**,"** *Education Week.*

Helfand, D. January 9, 2003. "State Keeps Education Standards," *Los Angeles Times.*

Henriques, D. B. September 2, 2003. "Rising Demands for Testing Push Limits of Its Accuracy," *The New York Times.*

Holland, R. December 10, 2003. "Some Schools Miss MCAS Progress Targets," *Brookline Tab.*

Howard, S. December 1, 2003. Quoted in Susan Ohanian, "Bush Flunks Schools," *Nation*.

Hutton, T. July 2003. "Critics Argue that NCLB's Adequate Yearly Progress Provisions Mislabel Schools," National School Boards Association. Available online at http://www.nsba.org/site/doc_cosa.asp?TRACKID=&VID=50&CID=1046&DID=31716

Jacques, N. January 31, 2004. "The Grades Are In," *Battle Creek Enquirer*.

Kane, T. J. and Staiger, D. O. April 2001. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." Paper presented at a Brookings Institution Conference.

Kane, T. J. and Staiger, D. O. March 2001. "Improving School Accountability Measures." Available online at http://papers.nber.org/papers/W8156.

Kingsbury, G. *et al.* November 24, 2003. "The State of Standards," Northwest Evaluation Association. Available online at http://www.young-roehr.com/nwea/.

Lee, J. April 7, 2004. "How Feasible is Adequate Yearly Progress (AYP)? Simulations of School AYP 'Uniform Averaging' and 'Safe Harbor' under the No Child Left Behind Act." *Education Policy Analysis Archives*, V. 12, N. 14. Available online at http://epaa.asu.edu/epaa/v12n14/.

Linn, R. L, Baker, E., Herman, J. Spring 2002. *The CRESST Line.* Available online at http://cresst96.cse.ucla.edu/products/newsletters/CLSpring02final.pdf.

Linn, R. L., Baker, E. L., and Betebenner, D. W. August/September 2002. "Accountability Systems: Implications of the No Child Left Behind Act of 2001." *Educational Researcher.*

Linn *et al.* Fall 2002. "Minimum Group Size for Measuring Adequate Yearly Progress," The CRESST Line. Available online at http://www.cse.ucla.edu/products/newsletters/CL2002fall.pdf.

Linn, R. L. September 1, 2003. "Performance Standards: Utility for Different Uses of Assessments," *Education Policy Analysis*

Archives, 11(31). Available online at http://epaa.asu.edu/epaa/v11n31/.

Linn, R. L. Winter 2003. "Requirements for Measuring Adequate Yearly Progress," Policy Brief 6, National Center for Research on Evaluation, Standards, and Student Testing.

MacDonald, C. November 30, 2003. "U.S. Flunks Top Metro Schools," *The Detroit News.*

Maine Education Association. March 2003. "ESEA Stinks," News and Views, *Online MEA*. Available online at http://www.maine.nea.org/dir2/esea_stinks.htm.

Marks, A. July 21, 2003. "As Schools 'Fail,' Parents Talk Transfers," *The Christian Science Monitor*.

Massachusetts Department of Education web site. 2003. Available online at http://www.doe.mass.edu/news/news.asp?id=1714.

MassINC. Fall 2003. The Center for Education Research and Policy, "No Child Left Behind: Opportunity or Obstacle for Massachusetts?" Available online at www.massinc.org/about/cerp/research/ NCLB%20Superintendent%20Survey.pdf.

Mathews, J. January 27, 2004. "What the Media Are Missing: Reports of Average Test Scores Mask Improvements Made by Minorities," *The Washington Post*.

Miller, K. September 10, 2003. "U.S. Education Law Nets Mishmash of Results," *Palm Beach Post*.

Montoya, B. October 10, 2003. "No Child Left Behind: Schools Fall Short of Goals," *Las Vegas Review-Journal*.

Pellegrino, J. W., Jones, L. R., and Mitchell, K.J. (Eds.). 1998. National Academy of Sciences, *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress.* Washington, DC: National Academy Press.

National Education Association. May 2003. "No Child Left Behind?" *NEA Today*.

Novak, J. R. and Fuller, B. December 2003. "Penalizing Diverse Schools? Similar Test Scores but Different Students Bring Federal Sanctions," *Policy Analysis for California Education (PACE)*, Policy Brief 03-4. Available online at http://pace.berkeley.edu/pace_publications.html.

Olson, L. May 5, 2004. "States Seek Federal OK for Revisions," *Education Week.*

Pardington, S. November 9, 2003. "Education Overhaul's Unintended Outcome," *Contra Costa Times.* Available online at http://www.bayarea.com/mld/cctimes/living/education/7220820.htm

Pinzur, M. I. August 8, 2003. "State Schools Fail to Meet New Federal Test Standards," *Miami Herald.*

Prah, P. December 9, 2002. "New Rules May Guarantee 'F's For Many Schools, *Stateline.org*.

Prah, P. June 16, 2003**. "**States Get Leeway to Meet Education Law**,"** *Stateline.org*.

Putnam, J. July 11, 2003. "Nearly All Troubled Schools in High-Poverty, Minority Neighborhoods," *Ann Arbor News.*

Rhoades, K. and Madaus, G. May 2003. "Errors in Standardized Tests: A Systemic Problem," National Board on Educational Testing and Public Policy. Available online at http://www.bc.edu/nbetpp.

Riley, C. November 16, 2003. "Academic Reward System to Spotlight State's Top Schools," *The Tennessean*.

Rodriguez, N. C. November 18, 2003. "470 Schools Fall Short on New Standards," *The Courier-Journal*.

School Funding Services. January 30, 2004. Available online at http://www.schoolfundingservices.org/news/hot_tips.01.30.04.pdf.

Simmons, T. July 30, 2003. "Law Threatens Some Schools," *The News & Observer.*

Stecher, B. *et al.* 2003. "Working Smarter to Leave No Child Behind: Practical Insights for School Leaders," RAND Corp.

Sunderman, G. L. and Kim, J. February 2004. "Large Mandates and Limited Resources: State Response to the No Child Left Behind Act and Implications for Accountability," Harvard Civil Rights Project.

Thomas, K. and DeBarros, A. August 5, 2002. "School 'Excellence' Thrown a Grading Curve," *USA Today*.

Tofig, D. August 7, 2003. "'Listed' Schools: Some See Stigma in 3Numbers Game," *Atlanta Journal Constitution*.

Truman, C. November 19, 2003. "What the School Scores Really Tell," *Lexington Herald Leader*.

Vaishnav, A. December 5, 2003. "School Ratings Raise Concern for Hispanics," *The Boston Globe.*

Winerip, M. October 8, 2003. "How a Good School Can Fail on Paper," *The New York Times.*

Table 1-1: Data on Schools Not Making Adequate Yearly Progress (AYP)

Source: State Department of Education websites and press reports. States also periodically revise their lists. Because of variations in state terminology and data reporting, these numbers are not completely comparable across states.

STATE	# OF PUBLIC SCHOOLS (used to make AYP decisions)	# OF SCHOOLS NOT MAKING AYP AT LEAST ONE YEAR	% OF ALL PUBLIC SCHOOLS	# OF SCHOOLS IN SCHOOL IMPROVEMENT, CORRECTIVE ACTION, OR RESTRUCTURING (Not making AYP 2 or more years)	% OF ALL PUBLIC SCHOOLS	Updated
AL	1,547	71	4.6%	46**	2.97%	2/17
AK	488	282	57.8%	66	13.5%	2/4
AZ	1,695	403	23.8%	219	12.92%	2/17
AR	1,130	281	24.9%	17	1.5%	3/25
CA	8,710	3,947	45.3%	1,991	22.86%	2/24
CO	1,613	436	27.0%	82	5.08%	2/17
CT	1,079	230	21.3%	14	1.3%	3/23
DC	151	83	55.0%	15	10.1%	4/5
DE	171	97	56.7%	12	7.0%	3/9
FL	3,182	2,466	77.5%	45	1.41%	2/17
GA	1,999	730	36.5%	257	12.9%	
HI	280	168	60.0%	84	30.0%	4/9
ID	645	428	66.4%	43	6.67%	2/17
IL	3,919	1,688	43.1%	577	14.72%	2/17
IN	1,891	97*	5.13%	50*	2.64%	
IA	1,442	156	10.8%	11	0.76%	2/17
KS	1,413	175	12.4%	30	2.1%	3/3
KY	1,179	470	39.9%	25	2.12%	2/17
LA	1,375	620	45.1%	69	5.0%	3/25
ME	707	167	23.6%	10	1.41%	4/13
MD	1,403	518	36.9%	131	9.3%	3/5
MA	1,694	625	36.9%	208	12.3%	4/13
MI	3,472	896	25.8%	363	10.5%	3/5
MN	1,949	155	8.0%	38	2.0%	3/31
MS	870	220	25.3%	7	0.8%	3/5
MO	2,055	1,009	49.1%	32	1.56%	4/15
MT	858	212	24.7%	40	4.7%	3/5
NE	1,220	269	22.1%	6	0.5%	4/9
NV	517	221	42.8%	21	4.1%	4/7
NH	448	140	32.1%	11	2.5%	3/1
NJ	2,448	1,051	42.9%	264	10.8%	

Table 1-1: Continued

STATE	# OF PUBLIC SCHOOLS (used to make AYP decisions)	# OF SCHOOLS NOT MAKING AYP AT LEAST ONE YEAR	% OF ALL PUBLIC SCHOOLS	# OF SCHOOLS IN SCHOOL IMPROVEMENT, CORRECTIVE ACTION, OR RESTRUCTURING (Not making AYP 2 or more years)	% OF ALL PUBLIC SCHOOLS	Updated
NM	780	164	22.0%	95	12.2%	3/25
NY	4,186	1,047	25.0%	729	17.4%	4/5
NC	2,252	1,194	53.0%	36	1.6%	3/25
ND	497	47	9.5%	23	4.6%	3/9
OH	3,715	815	21.9%	191	5.1%	3/29
OK	1,796	405	22.3%	46	2.6%	4/13
OR	1,231	333	27.1%	7	0.57%	3/9
PA	2,768	1,091	39.4%	298	10.8%	3/25
RI	313	98	31.31%	27	8.6%	3/25
SC	1,072	817	76.21%	90	8.4%	3/31
SD	737	196	26.6%	32	4.3%	4/12
TN	1,650	746	45.2%	62	3.8%	4/15
TX	7,733	563	7.3%	9	0.12%	3/9
UT	865	246	28.4%	6	0.7%	3/9
VT	307	39	12.7%	9	2.9%	3/9
VA	1,806	740	41.0%	66	3.7%	3/25
WA	1,613	436	27.3%	50	3.1%	3/15
WV	728	295	40.5%	33	4.5%	3/11
WI	2,019	171	8.5%	68	3.4%	3/11
WY	364	55	15.1%	0	0%	3/8
US TOTAL	**87,982**	**27,712**	**31.50%**	**6,565**	**7.46%**	

* State only released list of schools not making AYP for two or more years, and did not release list of schools not making AYP for just one year.
** Title I schools only.

Prepared by National Education Association Great Public Schools Action Plan. Used with permission of NEA.

Table 1-2: State AYP Starting Points

State		Reading/ELA	Mathematics
Alaska		64.03	54.86
Arizona	Gr 3	44	32
	Gr 5	32	20
	Gr 8	31	7
	High	23	10
Arkansas	K-5	31.8	28.2
	Gr 6-8	18.1	15.3
	Gr 9-12	19	10.4
California	Gr 2-8	13.6	16
	High	11.2	9.6
Colorado	Elementary	77.5	79.5
	Middle	74.6	60.7
	High	80.3	50.5
Connecticut	Elementary	55	64
	Middle	55	64
	High	62	59
DC	Elementary	30.3	38.4
	High	13.7	19.8
Delaware		53.9	30
Florida		30.68	37.54
Georgia	Gr 3-8	60	50
	High	88	81
Hawaii		30	10
Idaho		66	51
Illinois		40	40
Indiana		59	57
Iowa	Elementary	65	64
	Middle	61	63
	High	69	69
Kansas	K-8	51.2	46.8
	High	44	29.1
Kentucky	Elementary	47.5	22.73
	Middle	45.6	16.51
	High	19.26	19.84
Louisiana		36.9	30.1
Maine	Elementary	34	12
	Middle	35	13
	High	44	11
Maryland	Gr 3	58.1	65
	Gr 5	65.7	55
	Gr 8	59.9	39.7
	High	61.4	43.4
Massachusetts		39.7	19.5
Michigan	Elementary	38	47
	Middle	31	31
	High	42	33
Minnesota	Gr 3	62.75	66.17
	Gr 5	69.89	65.35
	High		
Mississippi	Gr 3	61	72
	Gr 4	66	49
	Gr 5	58	35
	Gr 6	51	39
	Gr 7	36	19
	Gr 8	30	23
	Gr 10	16	Algebra I
Missouri		18.4	8.3
Nebraska	Gr 4	62	65
	Gr 8	61	58
	Gr 11	66	62
Nevada	Elementary	32.4	37.3
	Middle	37	38
	High	91	58
New Hampshire	Gr 3-8	60	70
	High	64	52
New Jersey	Gr 4	68	53
	Gr 8	58	39
	Gr 11	73	55
New York*	Elementary	123	136
	Middle	107	81
	High	142	132
North Carolina	Gr 3-8	68.9	74.6
	Gr 10	52	54.9
North Dakota	Gr 4	65.1	45.7
	Gr 8	61.4	33.3
	Gr 12	42.9	24.1
Ohio	Elementary	40.5	35.9
	Middle	36	36.8
	High	78	53.1
Oklahoma		62.2	64.8
Oregon		40	39
Pennsylvania		45	35
Rhode Island	Elementary	76.1	61.7
	Middle	68	46.1
	High	62.6	44.8
South Carolina		17.6	15.5
South Dakota	Elementary	65	45
	High	50	60
Tennessee	Elementary	77.1	72.4
	High	86	65.4
Texas		46.8	33.4
Utah	Gr 3-8	65	57
Vermont*		Grade Span	Start Pt.
	Math	2, 4	314
	ELA	2, 4	385
	Math	8	287
	ELA	8	342
	Math	2, 4, 8, 10	293
	ELA	2, 4, 8, 10	380
	Math	8, 10	277
	ELA	8, 10	339
	Math	10	268
	ELA	10	345
	Math	2, 4, 8	306
	ELA	2, 4, 8	381
Virginia		60.7	58.4
Washington	Gr 4	52.2	29.7
	Gr 7	30.1	17.3
	Gr 10	48.6	24.8
Wisconsin		61	37
Wyoming	Gr 4	30.4	23.8
	Gr 8	34.5	25.3
	Gr 11	48.4	35.8

Note: Prepared from data in state plans submitted to the U.S. DOE or obtained from state education departments between summer 2003 and spring 2004. Not all states have yet determined starting points for calculating AYP or we were unable to obtain information. Missing states had no data available. * VT and NY have calculated an index score rather than a percentage point. The DOE URL for state plans is http://www.ed.gov/admins/lead/account/stateplans03/index.html .

Table 1-3: State Cell Sizes: Minimum Number of Students to Be in a Cell

State	Min. # students to use subgroup in AYP	Min. # to calculate partici-pation (if separate)
Montana	95% CI, one tailed	
North Dakota	uses 99% CI	
Maryland	5 and test of statistical significance	
South Dakota	10 and a 99% CI	
Louisiana	10 with CI of 99% for AYP	40
Utah	10 and 99% CI	40
New Hampshire	11	40
Massachusetts (as of '04)	20 students/yr for 2 yr rating; 15 in any one year; SE=2.5 for >50; SE=4.5 for 20<X<50; won't report without 95% CI	
Minnesota	20	40
Alaska	20	
New Jersey	0 and 95% CI	
Maine	20, will combine up to 3 yrs to reach min sample size; uses 95% CI	41
Arkansas	25	
DC	25	
New Mexico	25	
Nevada	25, combine years to bring schools to min	
Oregon	42 scores (21-28 students)	
Kentucky	10 per grade with a min of 30	
Arizona	30	
Colorado	30	
Indiana	30	
Kansas	30	
Michigan	30	
Missouri	30	
Washington	30	
Hawaii	30	40
Iowa	30	40
Vermont	30 over 2 years, plus 99% CI; small school review	
Nebraska	30, 45 for SPED	
Florida	30, except "small schools" having population <30 but >10	
Oklahoma	30 for all and regular ed; 52 for subgroups (99% CI); data aggregated across years to reach minimum of 30; might combine small schools scores	40
Wyoming	30	40
Ohio	30 except 45 for SPED	40
Puerto Rico	30 except 45 for SPED	40
Idaho	34	
Alabama '02-'03	40	
Connecticut	40	
Delaware	40	
Georgia	40	
Illinois	40	
Mississippi	40	
New York	40	
N. Carolina	40	
Texas	40; 50 for sub-groups; small schools can average yrs	40
Pennsylvania	40 will combine years to meet min.	
S. Carolina	40, combine up to 3 yrs	
Wisconsin	40; except 50 for SPED	
Rhode Island	45	
Tennessee	45	
Virginia	50	
W. Virginia	50	
California	100 scores or 50 students	

Note to Table I-3:

CI = Confidence Interval
SE = Standard Error
SPED = Special Education/Disability

Chart 1-1: Cell Sizes