Is Your Child's School Effective?

Don't Rely On NCLB to Tell You

Checked: No Child Left Behind Act of 2002, Title I: Adequate Yearly Progress
Florida A+ Plan: School Grades

Checked by Paul E. Peterson and Martin R. West

No Child Left Behind (NCLB), the federal school-accountability law, is widely held to have accomplished one good thing: require states to publish test-score results in math and reading for each school in grades 3 through 8 and again in grade 10. The results appear to be telling parents whether their child's school is doing a better job than the one across town, in the neighboring city, or across the state.

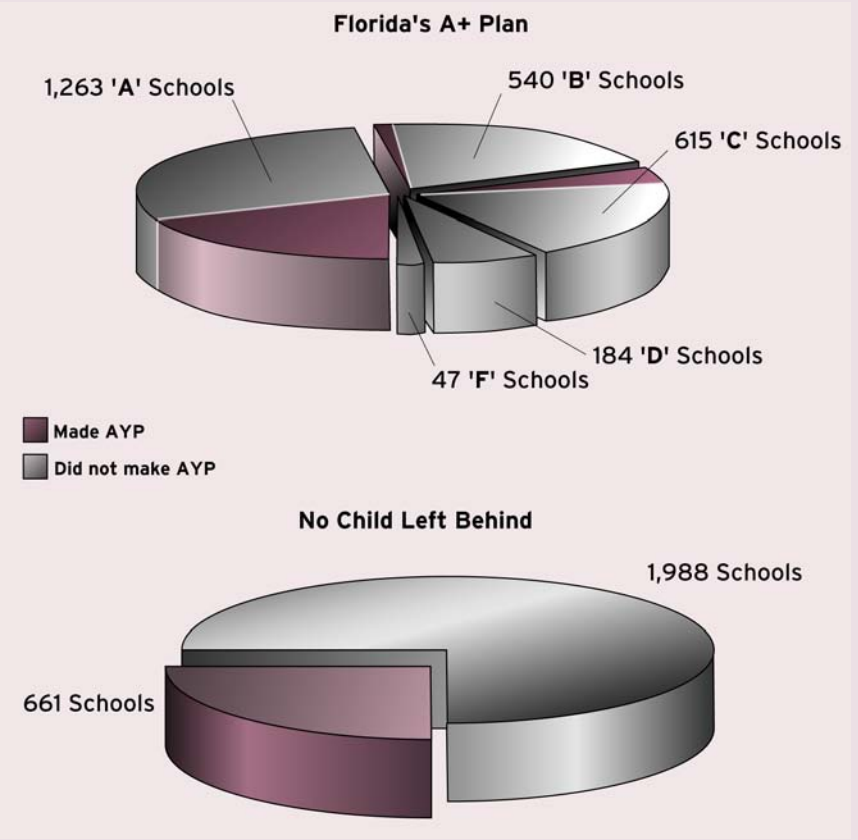
But accountability works only if the yardstick used to measure performance is reasonably accurate. Unfortunately, the yardstick required by the federal law is not. Our analysis of its workings in Florida reveals it to be badly flawed and not as accurate as the measuring stick employed by the state of Florida for similar purposes.

To her credit, Secretary of Education Margaret Spellings has apparently recognized the need to fix the NCLB yardstick. In November 2005, she announced a pilot program that would allow a few selected states to incorporate student growth into their AYP grading scheme. Although 20 states initially requested to participate, only 2—Tennessee and North Carolina—have so far been given the go-ahead, and the modifications they have been allowed to make are relatively minor. Meanwhile, the yardstick to be used by the other 48 remains as defective as ever.

Part of the problem is that NCLB makes only crude distinctions between schools achieving performance benchmarks and schools not doing so. Florida's grading system divides schools into five different categories, just as teachers do when they grade students on a scale from A to F. (See Figure 1 for the number of schools that received each mark.) Another part

Anatomy of Two Grading Systems (Figure 1)

In 2004 a quarter of all Florida schools made Adequate Yearly Progress (AYP) under No Child Left Behind (NCLB). Nearly half of Florida's schools earned an A under the state's own accountability system, but a majority of those schools did not make AYP.



Note: Grading based on school performance for 2003–04 school year. Only those schools subject to both state and federal accountability systems are included.

SOURCE: Florida Department of Education

of the problem is that the federal approach pays only a passing nod to the improvement made by individual students, while Florida's own method takes into account how much specific students have learned in a given year—exactly what parents care about.

It is not that Governor Jeb Bush (and his legislature) got it exactly right, while his brother (and Congress) ran amuck. But there is little doubt that NCLB needs repairing, something that Congress can do when the federal law is reauthorized.

Measuring Quality

Finding the right yardstick is no easy task. Not everyone agrees on what makes for a good school. Some reject test scores, while others care more about building students' character than boosting their academic achievement. But Congress took a clear stance on the issue in NCLB when it determined that a school with subpar student test scores in reading and math is not doing its job. Most Americans would agree that schools should aim to ensure that all students are proficient in these core subjects.

NCLB requires states to divide schools into those making "Adequate Yearly Progress" (AYP) toward the goal of having all of their students proficient in math and reading by 2014 and those that aren't. While the term "progress" would seem to imply that the law considers how much students are learning over time, the federal system in fact is based on a series of snapshots that fail to track individual students from one year to the next. Instead, to make AYP, schools must meet statewide targets for the percentage of students each year who are proficient. Those targets are gradually increased until they reach 100 percent in 2014. The percentage of proficient students within various subgroups, broken out by ethnicity, income, disability, and English-language-learner status, must also meet these same targets. If a school does not make AYP for two consecutive years, parents are given the choice of another school and, after five failing years the school is to be restructured.

But does the AYP yardstick actually distinguish between higher- and lower-quality schools? The answer to this question is best obtained by looking at how much students at the school know at the end of the year, as compared to how much those same students knew one year previously. If students are making large achievement gains, the school would seem to be more effective than if student improvement is meager or nonexistent.

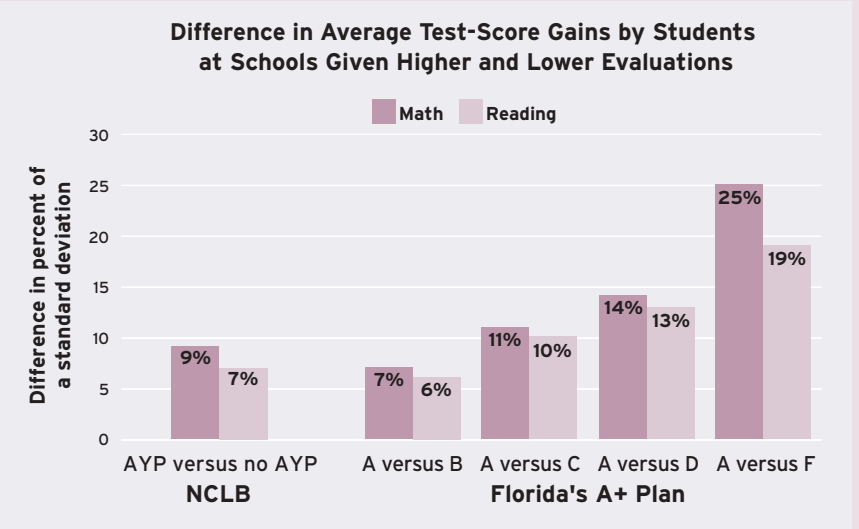
Surprisingly, in much of the United States, it is not possible to track an individual student's achievement over the course of a year to determine how well the federal yardstick identifies schools where students are learning the most. In Florida, however, the topic can be explored systematically because that state's Department of Education has assembled an impressive warehouse of data on student performance. As long as students remain within the state, it is possible to track how well most of them are doing from one year to the next on the Florida Comprehensive Achievement Test (FCAT), the exam the state uses to comply with NCLB requirements. (Privacy concerns preclude general release of the data, but qualified researchers who sign a confidentiality agreement can apply for access.)

Checking the Federal Yardstick

We drew on this information to calculate how much students learned, on average, in each school in Florida during the 2003–04 school year. We first subtracted from each student's test-score performance the child's demonstrated knowledge the previous year. We then adjusted those one-year-gain scores to take into account a statistical

What's in a Grade? (Figure 2)

NCLB's "either make AYP or not" accountability system does not perform as well as Florida's own A to F grading system in distinguishing schools where students are learning more from those where they are learning less.



SOURCE: Authors' calculations from 2003–04 Florida Department of Education data

To be wrong nearly three times out of ten does not inspire confidence—especially when one can get it right half the time simply by random guessing.

property that artificially generates larger gains for initially low-performing students (and smaller gains for high performers). Finally, we compared the average gains by students in schools meeting and not meeting the requirements for AYP.

The results were telling. On average, students in schools making the AYP target gained on their math achievement test an amount that was only 9 percent of a standard deviation more than the amount gained by students at schools said *not* to be making the AYP grade. The difference in gains in reading was just 7 percent of a standard deviation (see Figure 2). A full standard deviation's worth of progress equates to about four years of elementary schooling, so gains of 9 percent total a bit more than a third of a school year. A difference of that magnitude is surely worth noting, yet it is hardly enough to warrant saying one school is adequate while the other is not.

Nor is it the case that schools making AYP are those doing a much better job with minority student populations. In math, the differences in gains made by African Americans and Hispanics at AYP schools and non-AYP schools are 11 and 12 percent of a standard deviation, respectively. In reading, the difference in gains for both groups is 6 percent of a standard deviation. Clearly, such differences are not so dramatic as to be the basis for federal intervention.

Schools face varying challenges that depend in part on the populations they serve, so perhaps the federal yardstick does better when those challenges are considered. But that proved not to be the case. When we adjusted the gains made by students in each school to take into account a wide variety of individual and peer-group background characteristics, such as ethnicity, English-language-learner status, family income, and student mobility rates, the yardstick's performance actually worsened. In fact, the apparent benefit of attending a school that had made AYP was only 4 percent of a standard deviation in math performance and just 2 percent of a standard deviation in reading. To be credible, a grading system must do better than that.

Still another way of thinking about the accuracy of the NCLB yardstick is to calculate the probability that AYP identifies correctly the higher-performing of any two schools being compared. Of course, any two-category classification system will get it right 50 percent of the time, by chance alone, just as one can guess correctly half of the time which way a coin will flip.

How much better than chance did the NCLB grading system do in Florida in 2004? In math, a school that made AYP outperformed a random non-AYP school 71 percent of the time. In other words, 29 percent of the time the school in which students are making smaller gains is the one that passed AYP, a pretty hefty error rate (see Figure 3). In reading, that error rate was 28 percent. To be wrong nearly three times out of ten does not inspire confidence—especially when one can get it right half the time simply by random guessing.

So error-prone an emissions-testing program would soon invite the wrath of the auto-owning public.

Testing Florida's Approach

But can any other accountability system, especially one put together by a legislative body, do any better? Are we using the perfect to criticize the good? We can check this by comparing the federal yardstick with the one used by Florida as part of its state accountability system.

Florida's A+ Plan for Education (A+ Plan) rewards schools for ensuring that their students reach a minimum level of proficiency in math and reading, just as NCLB does. But unlike the federal grading system, the A+ Plan bases half of its points on the percentage of students in each school who improved their performance against state standards over the previous year. Equally important, it divides schools into five easily recognized categories that range from A to F, instead of just the two bureaucratically labeled categories employed by the federal government.

The Florida accountability system has its own limitations. But by having five categories, A through F, it provides parents and taxpayers with a good deal of useful information. Admittedly, some of the finer distinctions attempted by the Florida A+ grading scheme do little better than the federal AYP grading scheme. In 2004, for example, average learning gains in math were only 7 percent of a standard deviation higher in A schools than in those given a B (see Figure 2).

But the performance of the A+ Plan

improves when schools are assigned significantly different grades. The math learning gap between A and C schools was 11 percent; between A and D schools it was 14 percent, and the gap between A and F schools differed by 25 percent of a standard deviation. Put in more familiar language, the one-year difference between A and F schools amounted to more than a full year's worth of learning. In reading, the differences were almost as large.

As with AYP, we calculated an error rate for the Florida grading system, the chance that one would make a mistake—that is, pick a school where average learning rates were lower—if one picked a school solely on the basis of its official grade. Once again, the Florida A+ Plan can be seen to be employing a more accurate measuring stick than the NCLB one, where the error rate, it should be remembered, was nearly 30 percent. Under Florida's own accountability plan, parents would make an error 30 percent of the time if they chose an A school over a B school on that basis alone. But as Figure 3 shows, mistakes happen much less frequently if one picks an A school rather than a C, D, or F school. Indeed, one can have as much confidence in Florida's distinction between an A and an F school as the Food and Drug Administration requires when evaluating drugs subject to rigorous clinical trials.

The Florida system also does a better job of isolating the seriously defective schools, helping state and local officials identify exactly where attention is needed. In 2004, only 47 of the state's 2,649 schools were given an F, while 184 were given a D. Meanwhile, under the federal yardstick, 75 percent of schools did not make AYP, including more than half of the schools Florida had given an A (see Figure 1).

As these numbers suggest, having two accountability systems operating simultaneously has generated a great deal of confusion in Florida, as it has in other states. Things could be improved by melding both systems into one, but only if the revised system can do a better job of identifying schools where student achievement is rising and of isolating the worst-performing schools for remediation.

A National Problem

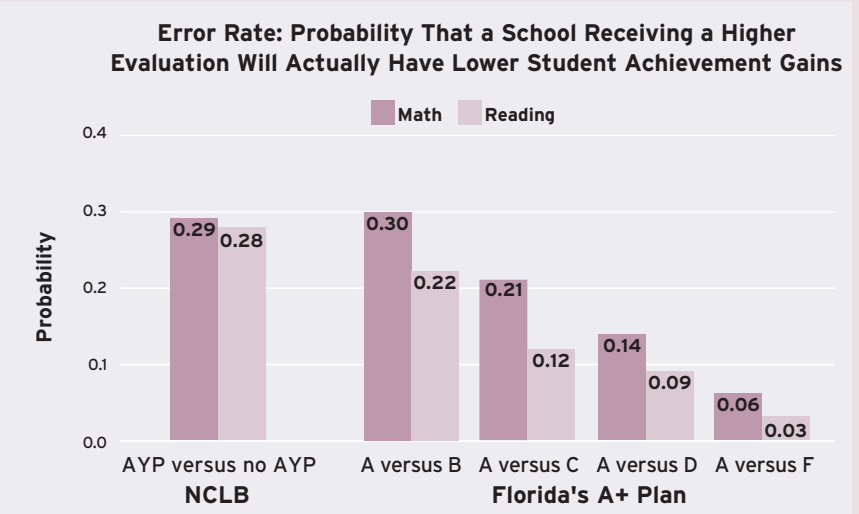
The shortcomings of the federal law's yardstick have a ready explanation. Because NCLB schools are evaluated primarily on the basis of achievement *levels*, the evaluation cannot readily detect how much *growth* is taking place within a school, simply because children come with dramatically different educational endowments. The correlation between school average levels and growth in the 2003–04 school year was just 0.63 in math and 0.71 in reading—a positive relationship, to be sure, but hardly one on which to construct a meaningful accountability system.

Some may argue that our focus on student growth is misplaced, that Congress, when devising its formula for gauging AYP, did not intend to distinguish good schools from less-effective ones. Its sole aim was to make sure that every school would by 2014 bring every student up to proficiency, and a level-based system is needed to direct reformers' attention to those schools and districts with the farthest distance to go.

But such claims are difficult to square with the legislators' designation of schools as not making "Adequate Yearly Progress," much less with the fact that the law gives families the option to attend another school if their

Error-Ridden (Figure 3)

When a school that is said to be making adequate yearly progress under NCLB is compared to one that is not, the chances are nearly 30 percent that students are learning less at the more highly rated school. This error rate under the Florida plan is just as high when A and B schools are compared, but it dwindles when A schools are compared to those given lower grades.



SOURCE: Authors' calculations from 2003–04 Florida Department of Education data

The Florida A+ Plan employs a more accurate measuring stick than does NCLB.

school twice fails to make AYP. Why let families move to another school without evidence that their children will learn more at the new address?

Of course, we have direct evidence about how the NCLB grading system is playing out from only one state. But scholars from the Northwest Evaluation Association have similarly documented the loose connection between growth scores and the level-based measures of school performance that underpin the AYP grading system in their database of 840 schools in 22 states, suggesting that the problem we have identified is hardly limited to Florida. Since the federal yardstick fails to zero in on how much each student is learning, it can hardly be otherwise.

It must also be admitted that most states could not have used growth scores when NCLB was enacted, simply because most states had not constructed the tracking system Florida has put together. Congress may have done all that it could in 2002. But since other states are now beginning to build their own warehouses of data that follow the progress of individual students, the time has arrived when a legislative fix should be feasible.

It will take Congress to do the job, since the original law was written with such specificity that it is virtually impossible to correct it through administrative action alone. Experienced authors know there's no such thing as good writing—only good rewriting. Let's hope that when NCLB is reauthorized Congress can avoid partisan bickering and use the information coming back from the states to improve on their first draft. People deserve to know that when the federal government says a school is not working, it means it.

Paul E. Peterson is professor of government at Harvard University and a senior fellow at the Hoover Institution. Martin R. West is an assistant professor at Brown University. Both serve as editors of Education Next.