FOREWORD

By Chester E. Finn, Jr., Michael J. Petrilli, and Amber M. Winkler

Way back in the 1990s, in that Mesozoic period known as the pre-No Child Left Behind (NCLB) era, most states were moving expeditiously to put K-12 accountability systems into place. These systems typically comprised academic content standards for the public schools and their pupils, regular assessments, school ratings, and, in some jurisdictions, the consequences that flowed from all of these.

The commonalities stopped there, however. Perhaps not surprising for America's much-touted "laboratories of democracy," several states made vastly different decisions about the specifics of their accountability systems. Academic standards in different locales were like night and day (as multiple Fordham analyses have shown), and in every way imaginable. Some were specific, others were vague. Some dealt with just the core subjects, others dived into art and music. Some were strong on knowledge, others concentrated on skills. Some embraced the teaching of evolution, others tiptoed around it. And on and on.

So, too, with state tests. Although most of these assessments were of the standardized, fill-in-the-bubbles-and-blanks variety, they varied in rigor and frequency, grade levels tested, and subjects examined. Some set high "cut scores," others low. Some reported performance against a single standard, others against multiple levels. And the school ratings that built on the results of said tests were a veritable (and literal) alphabet soup. A few states assigned letter grades to schools—sometimes A to F—based on the previous year's performance or, in some places, progress over time. Others developed complicated indices that pleased statisticians but befuddled parents and teachers. One state broke out data by race and income and only conferred laudatory labels on schools that served all groups of students well. Whether intended or not, experimentation was the name of the game.

But, regrettably, the let-a-thousand-flowers-bloom approach wasn't boosting mostly flatlined performance on the National Assessment (a.k.a. NAEP). Nor was it assuaging the widespread concern that America's competitive edge (perhaps like its youngsters?) was slowly dulling.

Enter NCLB. Its architects looked at this rocky landscape and saw chaos where others might have seen a healthy and diverse garden. They decided to bring uniformity to the country's uneven approach to K-12 accountability, though only in a few specific areas. States would still set their own standards, create their own tests, define proficiency however they liked, and determine their own rate of progress toward it. But all were now required to institute testing in reading and math annually in grades three through eight and once in high school, and all were expected to get 100% of their students to proficiency by 2014. They were also forbidden to deem schools as A-OK that garnered strong overall test results but failed to do the job for poor or minority or disabled students or kids with limited English proficiency. After all, NCLB was "an act to close the achievement gap," so accountability was bent to that gap-zapping purpose.

Consequently, when politicians and others say that they "agree with NCLB's goals," they ordinarily mean they accept the premise that good schools are those that serve all groups of students well, not just white or middle-class or high-achieving ones. In their view, besides shedding overdue sunshine on schools' actual performance with those groups, NCLB is exerting welcome pressure to make sure that none gets neglected.

So does that mean that today, thanks to NCLB, America has a common understanding of what makes for a successful school and how to spot a failing one?

Alas, no.

As this study shows, states are still singing different tunes when it comes to determining whether a given school is successful, or, in NCLB-speak, "makes adequate yearly progress."

The premise of this report is rather simple. Take a set of real schools, pretend that we can drag them around and

plop them down in various states, and see how many would make adequate yearly progress (AYP) in each place. If the United States had something akin to a shared notion of what it means to be a "good school" or a "bad school," we wouldn't see a huge variation from one jurisdiction to the next.

Yet what we found—as a handful of astute journalists and analysts have been finding out on their own—was something like the polar opposite. We discovered huge variation. In a few of the 28 states we studied, such as Wisconsin and Arizona, *almost all* of the elementary schools in our sample made AYP; in other states, such as Massachusetts and Nevada, *almost none* did. To put it colloquially, most of the schools in our sample would be considered failures in some states but just fine, even deserving of praise, in others. *These are the same exact schools, mind you.* Same students. Same teachers. Same achievement. What's different—sometimes drastically different—are the arcane rules that vary from state to state.

This report, written by our gifted and tireless colleagues at the Northwest Evaluation Association's (NWEA) Kingsbury Center, takes readers into the belly of the NCLB beast to understand how these variations came about. It builds on NWEA's groundbreaking work in Fordham's earlier *The Proficiency Illusion* study, which estimated the cut scores on reading and math tests in 26 states and concluded that NCLB's 100% proficiency requirement was encouraging a "walk to the middle" in terms of test rigor. But this study goes much farther, examining states' annual proficiency targets, minimum subgroup sizes, and confidence intervals—the mind-numbing details that yield wildly discrepant outcomes for individual schools.

Our purpose here is twofold. First, we want to bring greater transparency to the decisions that individual states have made in implementing NCLB. This stuff does get technical—we do our best in these pages to simplify wherever possible—and we suspect that there are many governors, legislators, education advocates, journalists, and school practitioners, not to mention parents and taxpayers, whose understanding of their state's approach to AYP is a bit hazy. Who could blame them? But with AYP determinations serving as life-or-death decisions for schools, it's critical that policy makers gain access to the "black box" that's driving these decisions. More than a few, we predict, will be surprised by how lax—or how rigorous—their state's AYP system is, relative to other states.

Second, we want to shine a spotlight on the maddening inconsistencies that riddle NCLB itself. We're surely not the first to note that it's snaring some good schools that deserve praise and letting some bad schools slip through its net. But we're not aware of any study that enables lay readers to examine the guts of this problem with such clarity.

Why, you may ask, is it a problem that verdicts vary so widely from state to state, when it comes to whether schools are making acceptable academic progress? Surely this variation existed before NCLB. Does it matter more today?

We think so, for three reasons. First, it surely demoralizes educators (and let's not forget students) to know that their own schools, deemed "in need of improvement" under NCLB, would be considered acceptable, perhaps even laudable, were they located in another locale. The capriciousness of NCLB breeds cynicism, which cuts against the idea of accountability itself—and certainly against efforts to revitalize truly bad schools and boost low-performing pupils.

Second, what drives the state-to-state variation in AYP results isn't a principled difference about what it means to be a good school. Instead, obscure, little-noticed, and ill-understood decisions around concepts like *cut scores, annual measurable objectives, minimum n sizes*, and *confidence intervals* are creating discrepant outcomes. We'd actually prefer it if the variations were based on things that truly matter, like whether schools are judged for their progress over time instead of for the previous year's performance, whether schools are helping all students make gains versus just those below a fixed level of proficiency, whether determinations hinge solely on reading and math or include such other core subjects as science and history, and so forth. Those would be legitimate reasons for discrepancy, issues worth arguing about—and

maybe welcoming divergent decisions from state to state. But that's not what we're seeing here. Without impugning the motives of state officials who made these decisions—especially since a case can be made that NCLB itself incentivized them to cut some corners and manipulate some rules to their schools' advantage—we are dismayed that such big differences emerge from such low-visibility selections among alternative paths.

Let's be clear, though, when it comes to AYP systems, harder isn't always better. We feel for states with high standards and rigorous tests that watch with horror as good schools get snagged as needing improvement because their special education or limited English proficient students aren't reaching targets. These states face a choice: either label virtually all their schools as failures, or tinker like crazy with minimum *n* sizes and confidence intervals and annual targets and all the rest. So we witness another unintended consequence of NCLB. Just as its call for "universal proficiency" encourages states to keep their cut scores low, so does its call to hold schools accountable for every single subgroup—including those with learning disabilities and limited English skills—encourage states to play around with the mechanics of AYP.

Third, the mere existence and promises of NCLB itself create the impression of a national accountability system. State variation around school ratings was fine when states also got to decide the penalties for schools not making the grade. But now every state labors under a rigid, federally prescribed, and inviolable cascade of interventions in low-performing schools. States are told in which year (of a school's not making AYP) to intervene in which way. The man in the street surely believes that it's a uniform accountability system. Yet it's not. All those sanctions and interventions, uniform though they are, are triggered by AYP systems that couldn't be more different. At best, there's a disconnect. At worst, it's complete chaos.

So what to do? Some politicians imply that NCLB might be "repealed." Not likely. NCLB is the umpteenth reiteration of the Elementary and Secondary Education Act of 1965, the vehicle through which most federal aid to K-12 education flows. Nobody is going to scrap it. The real issue, going forward, is what strings and conditions will be attached to those federal dollars in the name of accountability.

Another alternative is to tighten the screws by making states justify their decisions around *n* sizes and confidence intervals and so forth. That's what new Title I regulations, released in October by the Bush Administration, will require. They might help on the margins, but we're not optimistic.

One bold option would be to nationalize and standardize everything. Perhaps that's not as unthinkable as it once was, now that Washington is running large swaths of our economy. We could move to national standards, national tests, and a national definition of AYP. The Department of Education would determine each year which of the country's 100,000 public schools makes the grade.

But that's not what we'd recommend. Far from it. For it would push Uncle Sam deeper still into the hopeless morass of running schools and trying to turn around those that fail. And if there's anything that NCLB has taught us, it's that (1) the federal government doesn't have any better ideas about overhauling failing institutions than anyone else and (2) it can't ensure the ideas that it does put out there are well implemented and enforced. (We can only hope it knows more about turning around banks.)

We picture an altogether different approach to NCLB 2.0. Create incentives for states to sign on to common national standards and tests, through a process like the one being launched by the Council of Chief State School Officers, the National Governors Association, and Achieve. Ensure that the common assessments are rigorous and comprehensive. Publish the results of those annual tests for every school in the country, sliced every which way—by race/ethnicity, income, disability status, progress over time, and so on. And then stop.

That's right, stop.

Go back to the pre-NCLB world where each state gets to decide how to interpret those test results and what to do

about schools whose results don't satisfy it. Some places will likely return to grading their schools on an A–F curve. Others will obsess over student growth. Others will decide that including English language learners when calculating a school's rating doesn't make much sense. Let the states again differ in these and other ways. Civil rights groups and others that don't like state decisions can create their own school ratings, using the same uniform national data, accessible and transparent to all. So, too, could private organizations such as GreatSchools.net. We could reopen the debate about what it means to be a good school or a bad one. And then it would be up to the states to do something (or yes, nothing) about the schools that aren't making the grade.

We understand that this approach would move away from the ambitious, even utopian, rhetoric of the NCLB era. It would amount to admitting that the federal government actually cannot ensure that every child in America gets a world-class education. But what this strategy would do is ensure greater transparency around student achievement results—something this report shows is hard to come by—based on assessments that are rigorous and credible. And it would reinforce the idea that the states are still responsible for K-12 education and must make decisions in that realm that their own citizens will accept. Best of all, it would end the gamesmanship that has characterized the federal–state relationship for the past seven years.

This big study was the product of many hands and heads. At NWEA's Kingsbury Center, John Cronin and Michael Dahlin were the chief analysts and writers of the report. In addition to their first-rate analytical skills and attention to detail, they are a pleasure to work with. Special thanks go to the Joyce Foundation, and to our sister organization, the Thomas B. Fordham Foundation, both of which furnished funding for this and *The Proficiency Illusion*. Andrew Porter at the University of Pennsylvania and Martin West at Brown University provided expert feedback on methodology. René Howard and Christina Thomas painstakingly copyedited every word, figure, and table. Emilia Ryan created the sharp design. Here at Fordham, interns Molly Kennedy, Hannah Miller, Charlotte Underwood, Yusi Zheng, and Katie Wilczak and Fordham Fellow Ben Hoffman offered a multitude of assistance. Amy Fagan and Laura Pohl capably handled dissemination, and program associate Christina Hentges brought it across the finish line. We heartily thank them all.