THE Accountability Illusion

FEBRUARY 2009

By
John Cronin
Michael Dahlin
Yun Xiang
Donna McCahon

Foreword by
Chester E. Finn, Jr.
Michael J. Petrilli
Amber M. Winkler

THOMAS B. FORDHAM INSTITUTE
Advancing Educational Excellence

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

THE Accountability Illusion

By
John Cronin
Michael Dahlin
Yun Xiang
Donna McCahon

Foreword by
Chester E. Finn, Jr.
Michael J. Petrilli
Amber M. Winkler

FEBRUARY 2009

THOMAS B. FORDHAM INSTITUTE
Advancing Educational Excellence

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

TABLE OF CONTENTS

Executive Summary 3

Foreword 7

Preface 11

Introduction 15

Methodology 18

Findings 21

Limitations 45

Discussion 47

Appendices A, B, C 49

References 62

EXECUTIVE SUMMARY

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all students in grades three through eight achieve proficiency in reading and math by 2014, with a particular focus on groups that have traditionally been left behind. Under NCLB, states submit accountability plans to the U.S. Department of Education detailing the rules and policies to be used in tracking the adequate yearly progress (AYP) of schools toward these ambitious goals.

This study examines the NCLB accountability systems and the basic AYP rules for 28 states as they operate in practice. We did this by selecting 36 real schools from around the nation (half elementary, half middle)—schools that vary by size, achievement, diversity, and so on—and determining which of them would or would not make AYP when evaluated under each state's accountability rules.[1] In other words, if a particular school that made AYP in Washington were relocated to North Dakota, or Ohio, or Texas, would that same school also make AYP there? And if not, what factors within NCLB, and its implementation by the various states, explain this? Based on this analysis, what can we learn about how AYP determinations vary across the country—and, at least by inference, about the effectiveness of NCLB in ensuring that *all* students attain proficiency?

NCLB imposes strict expectations for schools—100% of their students must achieve proficiency by 2014—but gives states wide latitude in terms of key variables. Under the act, states have leeway to:

1. Craft their own academic standards, select their own tests, and define proficiency in reading and math as they like; as a result, proficiency standards (which take the form of cut scores[2] on state tests) vary widely in their rigor and consistency.

2. Establish their own annual targets (also called annual measurable objectives or AMOs) for moving students to the proficient level by 2014. Some states require schools to follow a linear trajectory to the 100% proficiency goal, seeking similar gains each year; others use a back-loaded trajectory (meaning that little improvement is required during the early years and much is required during latter years) to achieve this result.

3. Apply confidence intervals, or margins of statistical error, to schools' proficiency rates. When states use such intervals, it means that the percentage of students required to reach proficiency can actually be lower than the stated target. States also determine the confidence interval's size and how it is used.

4. Determine when the size of a student subgroup within a school is large enough that it must meet AYP targets. In other words, states decide whether particular subgroups of minority, low-income, or limited English proficient (LEP) students, for instance, are large enough that their test results must be counted separately for determining their school's AYP status, in addition to being counted within the general school population.

How do these multiple allowances for state discretion and variation affect AYP determinations from state to state? To find out, we evaluated the performance of students in 18 elementary schools and 18 middle schools relative to each state's proficiency cut scores and 2008 annual targets. We also applied confidence intervals to results, according to each state's rules, and evaluated the performance of all subgroups within a school that met or exceeded each state's minimum pupil-count requirement. This allowed us to estimate whether a school would meet most of the requirements needed to make AYP.

[1] We did not examine the impact of NCLB's "safe harbor" provision or other indicators such as attendance and test-participation rates. Nor were we able to consider the impact of the U.S. Department of Education's recent growth model pilot program, which allows states to track individual student achievement over time. We used school data and proficiency cut score estimates from academic year 2005–2006 and applied them against state AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

[2] A cut score is the minimum score a student must receive on the applicable state test in order to be considered proficient under that state's accountability system.

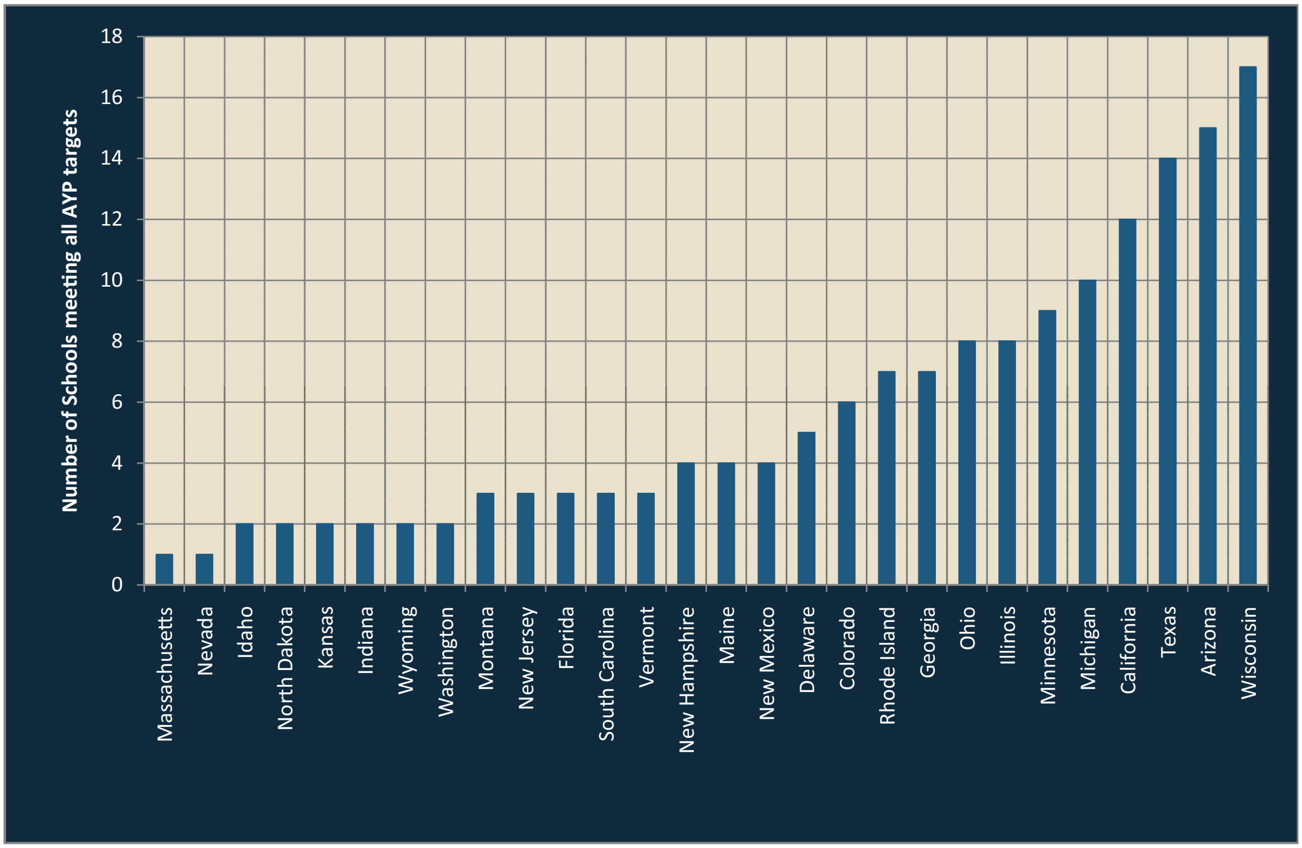

Figure ES-1. Number of sampled elementary schools that made AYP in 2008, by state

Here are the study's key findings:

- **Within the elementary school sample, the number of schools that made AYP varied greatly by state.** Almost all our sampled schools **failed to make AYP** in some states, and nearly all of these **same schools made AYP** in others. In Massachusetts, for example, a state with high proficiency cut scores and relatively challenging annual targets and AYP rules, **only 1 of 18** elementary schools made AYP; in Wisconsin 17 schools made AYP (Figure ES-1). Same kids, same academic performance, same schools—different states, different cut scores, different rules. And very different results.

- **There is more consistency across states with the middle school sample because so few of these schools made AYP in any states.** In 21 of the 26 states studied,[3] two or fewer middle schools made AYP. In no state did even half of the 18 middle schools meet the 2008 AYP requirements. This is mostly because the larger size of middle schools generally means that they have plenty of students with disabilities (SWDs) and minority, low-income,[4] and LEP pupils who are counted separately for accountability purposes. Although subgroups of minority students within our sample schools performed well enough to meet their annual targets in many states, almost all schools with a qualifying LEP or SWD subgroup failed to meet the targets for these groups in nearly every state.

- **When it comes to whether the performance of a subgroup will hurt a school's chances of making AYP, the state's decision relative to minimum subgroup size (called "*n* size")** is critical. Consider Chaucer Middle School, for example, the highest performing middle

[3] Two states (Texas and New Jersey) are not included in the middle school analysis because 8th grade cut scores were not available.

[4] Low-income students are those who receive a free or reduced-price lunch.

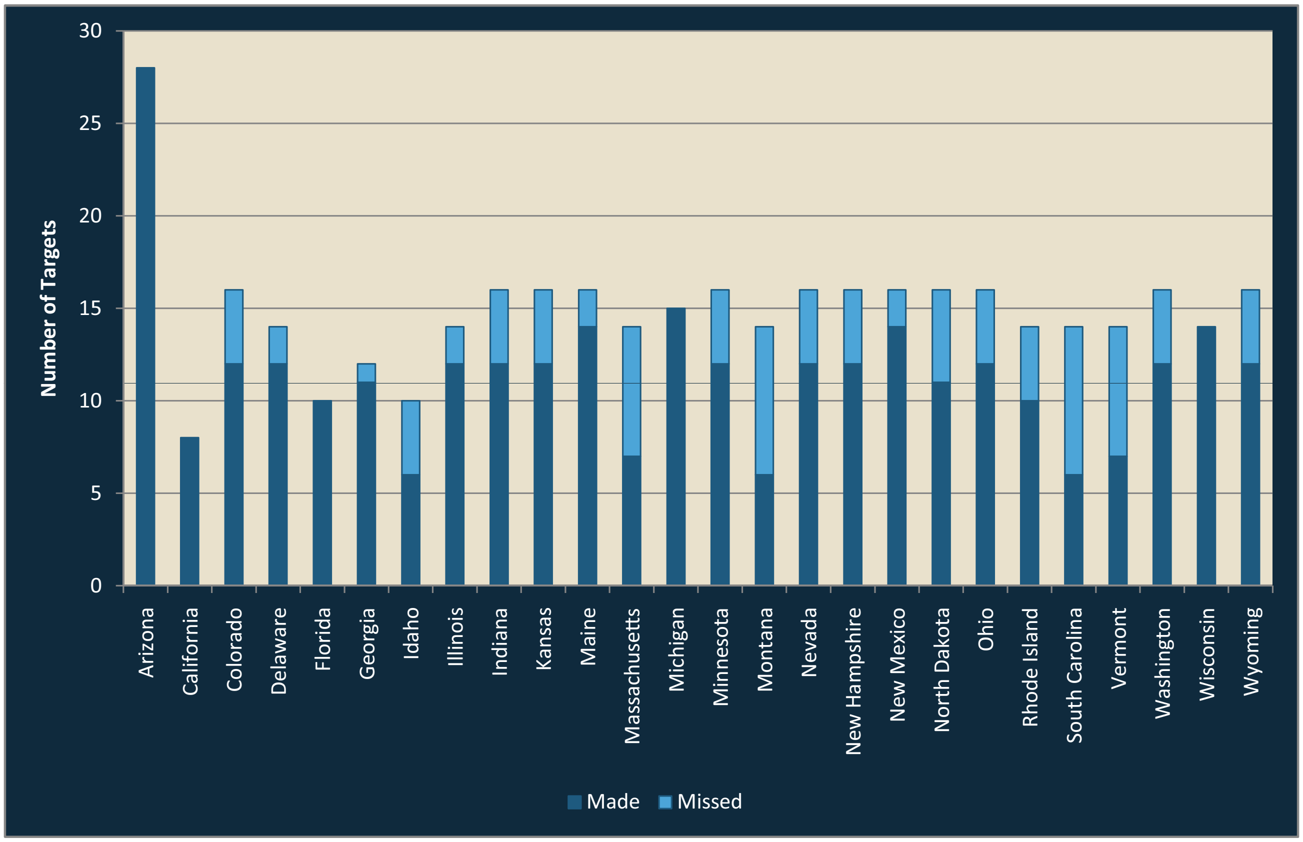

Figure ES-2. Number of subgroup targets met by Chaucer Middle School in 2008, by state

Note: Arizona has more targets because each grade level is considered a group unto itself. For instance, a middle school in Arizona with three grades and four subgroups has 3 × 4 × 2 (subjects) or 24 targets.

school in our sample (see Figure ES-2). Though it achieved strong performance overall and added greater value to its students' performance over time than most other schools in the country (and virtually all schools in the sample), it failed to make AYP in 21 of the 26 states because of the performance of its subgroups (if even one target is missed, as indicated by the light blue bars, the school does not make AYP in that state). In the states with relatively small *n* sizes, where Chaucer is held accountable for numerous subgroups (e.g., Nevada, New Hampshire, and North Dakota), it did not make AYP.[5] On the other hand, in states with large *n* sizes, where Chaucer is held accountable for fewer subgroups (e.g., Florida and California), it made AYP. Generally, the lower the state's *n* size, the more subgroups for which the typical school is accountable, and the more separate targets that school must hit.

Implications

For an accountability system to be effective, educators must believe that it is fair, consistent, and understandable. Unfortunately, the way NCLB rates schools appears to be idiosyncratic—even random—and opaque. Schools that make AYP in one state fail to make AYP in another. Those that are considered failures in one part of the country are deemed to be doing fine in another. Although schools are being told that they need to improve student achievement in order to make AYP under the law, the truth is that many would fare better if they were just allowed to move across state lines.

One of the adages of the NCLB era is that a child's zip code shouldn't determine her life chances. Indeed. But neither should a school's zip code determine whether or

[5] Arizona is an exception, but the number of subgroups in Arizona is large primarily because they treat each grade level as a subgroup. Grade levels are not subgroups in the same sense as low-income students, or LEP students would be considered a subgroup because they have no defining achievement related characteristic that distinguishes them from others.

not it makes AYP. Yet regrettably it often does. And so the success or failure of a given school under NCLB is driven as much by the way the law is implemented by its home state as it is by the performance of its students and the amount of progress they've made over the course of a year.

This is the Accountability Illusion.

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.

PREFACE

By John Cronin and Michael Dahlin

Set standards. Test students. Sanction schools that don't measure up.

This is the NCLB formula for accountability, and it seems simple and compelling. Thanks to the passage of NCLB, we have proficiency standards and testing for all students in grades through 3 through 8, plus one high school grade. We have a no-excuses requirement that 100% of students achieve these proficiency standards, and a firm deadline for achieving them by 2014. There are also strict sanctions imposed on schools that do not meet the Annual Measured Objectives (AMOs), the proficiency rates required to stay on track for the 2014 deadline.

This is NCLB's sixth year of implementation. Large numbers of schools have been identified as underperforming and many of those schools have been sanctioned. As far back as 2005, over 10,000 schools across the United States had failed to make adequate yearly progress (or AYP) for two years in a row, thus putting them in "program improvement" (National Education Association 2006). And this year, California alone has 2,241 schools, about 22%, in program improvement (*San Francisco Chronicle* 2008). These numbers have increased dramatically in the past three years and the pace will likely accelerate as the Act's 2014 deadline draws closer.

We have standards, we have deadlines, and now we have a large round-up of K-12 suspects. Were we as cynical as Captain Renault from the film *Casablanca*, a round-up of the usual suspects would be all we needed to maintain an illusion of accountability, and it would little matter whether our suspect schools were really culprits in some crime against learning. To their credit, Former President Bush, Senator Ted Kennedy, Margaret Spellings and others who have driven support for this reform are not Captain Renault. The 2007 blueprint for reauthorizing NCLB stated the sentiments of those who support NCLB in plain, ambitious terms; its goal being to deliver "…steady academic gains until all students can read and do math at above grade level, closing for good the nation's achievement gap between disadvantaged and minority students and their peers (pg. 1)" (U.S. Department of Education 2007). The statement is quite sweeping; it does not suggest that the law's intent is merely limited to eliminating achievement gaps within a state. Rather, her statement refers to these as *national objectives*, which can be achieved only by wiping out differences in the performance of groups of students across states.

The strategy for achieving these objectives under NCLB might be best described as a "strict-loose" approach. NCLB's requirements for setting standards, testing students, and specifying deadlines are clearly strict. However, NCLB is loose in giving states wide latitude to determine both the difficulty of the proficiency standards (or cut scores) and the annual benchmarks that schools must achieve in order to make "Adequate Yearly Progress" (AYP) between now and 2014. Furthermore, NCLB allows states to set their own accounting rules for how students are categorized for evaluation. These rules include, among others, determining the minimum number of students in various groups that are separately accountable under NCLB, whether to apply a confidence interval (or margin of error) to proficiency results and, if a confidence interval is applied, its size.

If educational equity is the goal, then the strict-loose approach must achieve some degree of consistency in results for it to be reached. After all, if we accept that a school ruled "in need of improvement" in Florida, would not get that same label if it happened to be in New Jersey, California, or Illinois, then we are not truly eliminating achievement gaps – we are merely replacing gaps based on race or poverty with gaps based on geography.

If the goal of ensuring that all students achieve high standards is a **national objective**, then it is reasonable to ask whether this "strict-loose" approach is producing some modicum of consistency. Thus we, alongside our colleagues from Fordham, undertook a study to investigate two research questions.

1. Is there enough consistency among the various state proficiency standards and objectives to conclude that expectations across the states are similar? Does making AYP reflect equivalent achievement across the various states?

2. Do states apply the standards, timelines, and the various state rules in a manner that results in consistent judgments about schools across states? Would a school that meets Florida's expectations, in reality, also meet the expectations of New Jersey, or California, or Illinois?

To investigate these questions, we found a sample of 36 schools that reflect the diversity within the American educational system. Students in these schools took achievement tests that predict their proficiency status on 28 state tests with a high degree of accuracy. From this achievement information, estimates of the school's proficiency rates could be produced for each of the states studied. Thus, if a school achieved a proficiency rate of 70% in Illinois, it was possible to estimate what that proficiency rate would be if the school were located in Wisconsin, Minnesota, Indiana or other states. Once the proficiency rate is known, we can determine whether that proficiency rate would have been sufficient to reach the state's annual proficiency targets (AMOs) and whether the school would likely make AYP. Finally, it's possible to estimate whether a school that made AYP in Illinois would also it in other states.

With respect to the first question, the results of this study demonstrate that proficiency standards across states are vastly different. Case in point: one elementary school in our sample that achieved a predicted 80% proficiency rate under Wisconsin standards, achieved a 52% proficiency rate under Massachusetts standards, and only a 19% proficiency rate in California.

But standards are only one part of the equation. Each state also has AMOs, which are timetables of targets that require increasing proportions of students to achieve proficiency between now and 2014 (the NCLB deadline for achieving 100% proficiency). This study and others (e.g., Chudowsky and Chudowsky 2008) show that these timetables vary as much as the standards. But what is the result?

Consider Wolf Creek Elementary, a California school in our sample. Its students achieved a 54% reading proficiency rate and met their AMO. If Wolf Creek were relocated to South Carolina, we estimate their students would achieve about the same proficiency rate, 53%, since South Carolina's reading cut scores are roughly comparable to California's. But this rate of proficiency would fail to meet South Carolina's AMO (hence Wolf Creek fails to make AYP). In other words, we could have the same students produce the same proficiency rate in two states, and get two very different AYP outcomes. To make matters worse, consider what happens if Wolf Creek is relocated to New Jersey (whose state test is easier to pass). The school's estimated proficiency rate now rises to 80%, but in New Jersey, 80% is not high enough to meet the AMO. But had we dropped Wolf Creek into Michigan, whose state test is roughly equal in difficulty to New Jersey's, 80% proficiency would have been high enough to meet the AMO. So in Michigan, Wolf Creek Elementary would make AYP! Does this seem confusing? Take heart, because it is!

Is Wolf Creek on the path to "all students achieving high standards"? Who knows? How could one possibly tell? Performances that were a hit in Fresno bombed in Trenton. A school we might call a rose in Ann Arbor would not smell as sweet in Spartanburg…

Of course we recognize that the background and achievement of students vary from state to state. But there's no reason to believe that there's less need for math and reading competence in California than there is in South Carolina. And even if NCLB is successful in getting 100% of students to proficiency by 2014, all it will mean is that we have created an Orwellian system in which all students are proficient, but some are more proficient than others.

The second question we asked in this study was whether the state accountability systems created under NCLB make consistent judgments about schools across the various states. Whether sanctions achieve their desired end depends on how effectively they are deployed. For the system to work, sanctions must target schools that are actually underperforming. Unfortunately, this study found little consistency across states in how NCLB is implemented, and rarely were adequately performing schools differentiated from underperforming ones.

Many years ago, one of the study authors taught high school. At this school, it was typical for nearly all the students enrolled in choir classes to receive "A" grades. One wouldn't know from the grading system that some of the students were highly-motivated, vocally gifted stars; that others were recreational singers of average talent; and that yet others took the class to get an easy grade. In this same school was another teacher who dedicated her efforts to finding *failure* somewhere inside every student. This teacher was legendary for giving pop quizzes, counting them triple if the students performed poorly, or discounting them by half if students performed too well.

In this study, state accountability systems fit both of these archetypes. Despite their large differences in achievement and growth, nearly all of the sample elementary schools made AYP under some accountability systems. Roughly one-third of the states have a combination of proficiency standards, AMOs, and rules that were met by the overall school populations in every single school within our sample. In such states, one could reasonably argue that students would be better served by higher proficiency standards, more aggressive targets, stricter rules, or perhaps all three.

On the other hand, many of the state accountability systems seemed designed to ensure school failure. Shockingly, the highest performing elementary school in our sample failed to make AYP in thirteen of the twenty-eight states studied, and the highest performing middle school failed in twenty-three states. Under the accountability systems in Massachusetts and Idaho, to cite two examples, every single middle school within our sample failed to make AYP.

The accounting rules used to define subgroups differ across states, and this one factor largely explains the indiscriminate effect of NCLB in certain states. NCLB requires that proficiency be achieved on the same timetable for all subgroups within a school, a goal meant to eliminate racial or income-based educational disparities. This "no-excuses" aspect is one of NCLB's most attractive features; it does not permit educators to write off the performance of minority or other traditionally disadvantaged groups. To the extent that NCLB has focused attention on improving the performance of these subgroups, it can be called a success.

While disaggregation is laudable, in practice the subgroup requirements cause the most diverse schools—particularly in states with more ambitious proficiency cut scores— to fail AYP. In about 30% (elementary sample) to 50% (middle school sample) of cases, low-income students failed to make their 2008 annual targets. In over one-half of the cases, one or more groups of minority students failed to make their AMO.

The results for limited English proficient (LEP) students and students with disabilities (those with Individualized Education Plans) were more depressing. These groups almost universally failed to meet AMOs regardless of the state they were in. In only 2% to 4% of the cases we evaluated did a group of LEP students actually achieve their AMO, even in states with relatively low proficiency cut scores and in states that "boost" their observed performance rates by reporting confidence intervals (or margins of error). Similarly, in only 2% to 6% of cases did students-with-disabilities (SWDs) achieve their targets. Ultimately even the highest performing schools—schools whose own LEP or SWD subgroups outperformed most or all of the same students in other schools—generally failed their AMO.[1]

Looking at the data, we would conclude that states have two possible strategies to cope with this problem, both of which are untenable. One is to avoid having subgroups. In general, schools within our sample that did not have LEP or SWD subgroups

[1] For reasons explained in the report, however, our estimates of SWD and LEP subgroup performance may be lower than is actually the case.

had a fighting chance of making AYP. So, if states were to set the minimum n size requirement so high that these subgroups escaped separate reporting, schools could up their AYP odds. The other solution would be to create proficiency standards so low that they could be met by 100% of students. Clearly, both of these solutions are at odds with NCLB's intended goals.

Simply put, it's a hard knocks life for states trying to implement NCLB in a manner consistent with its intent. When states adopt high standards, when they set AMOs on a rigorous timetable, when they establish rules about minimum subgroup sizes that are reasonable, then their schools are inevitably seen as failures under NCLB. For the schools in our sample, this was a plain, irrefutable fact. When confronted with these odds, educators in some of our better schools might be forgiven for feeling like new recruits in military basic training: They can make up their bunks immaculately, shine their boots to a high polish, learn all the drills to perfection, but still get 500 push-ups from the drill sergeant because he found a stray bristle on a toothbrush.

As currently implemented, NCLB is not a discriminating system. A tremendous amount of money and energy has been spent to create the impression that there is accountability, and there are large numbers of schools throughout the United States that are in some phase of sanctions. But the accountability is not coherent. We found states where most schools failed to make AYP and others where nearly every school made it. We found demonstrably good schools that failed to make AYP far too often, and some pretty mediocre ones that slide by in some states. Thus what seems like accountability is an illusion. Good schools get sanctioned, bad schools get off, and ultimately students get shafted, since maintaining this illusion has a cost. When good schools get sanctioned, resources are wasted and we risk causing quick-fix, panic driven, counterproductive change in schools that may ultimately hurt students. When bad schools get off, their students are denied opportunities (what we unfortunately now call "sanctions") that might lead to a better education, including the chance to attend a different school, or receive supplemental services, or simply obtain assurance that the workings of a perennially dysfunctional school will be addressed and corrected.

It 's long past time to dispel the accountability illusion.

INTRODUCTION

NCLB's accountability and intervention provisions were intended to identify and correct underperforming schools. The ultimate goal—for all students to reach high standards—will not be met if schools are graded inconsistently, yet it's well known that NCLB does not establish a uniform benchmark for determining whether schools make Adequate Yearly Progress (AYP), but, instead, allows for quite a bit of state discretion.

First, states can define proficiency in reading/English language arts (hereafter called reading) and math; as a result, proficiency standards vary widely in their rigor and consistency (National Center for Education Statistics 2007; Cronin, Dahlin, Adkins, & Kingsbury 2007a; Kingsbury, Olson, Cronin, Hauser, & Houser 2003). Second, NCLB allows states to establish their own timetables, or annual measurable objectives (AMOs) for moving all students to the proficient level by 2014. Some states require schools to follow a linear trajectory to the 100% proficiency goal, while others use "stair steps" or a back-loaded trajectory (i.e., more of the required improvement must be made in the final few years). Third, in an effort to recognize the potential for error in any assessment, NCLB permits states to use confidence intervals (a.k.a. margins of statistical error) in determining proficiency rates, and also allows states to define both the methodology for estimating the confidence interval and its size. Fourth, NCLB allows states to establish their own rules governing the size that a subgroup—such as Hispanic/Latino or low-income students—must attain within a school for the group's performance to be included in the school's AYP determination. States are allowed to determine the minimum size of these subgroups and, if the number of students in the group falls below this number, they are not counted separately as a subgroup for accountability purposes (though they are, of course, counted in the overall student population).

Given the various state interpretations of NCLB, it is reasonable to ask whether differences in standards, timelines, and rules lead to differences in the schools identified as ineffective. For example, if a school that made AYP in Washington were suddenly dropped into North Dakota, or Ohio, or Florida, or Texas, would it also make AYP there? And if not, what factors within NCLB explain this? Based on this analysis, what can we learn about the variation of the AYP systems used throughout the country? To explore these questions, this study looked closely at a group of 36 schools (18 elementary and 18 middle schools). The performance of these schools on a common assessment was used to estimate whether each school would have made AYP in each of the 28 states whose accountability systems were studied. In other words, this study examines how each school would fare if the 28 different standards and rules used to govern AYP decisions under the No Child Left Behind act (NCLB) in these 28 states were applied to them.

Literature Review

Whether a school makes AYP or not depends on many factors. In this particular study we focused on four of them. They are:

1. The difficulty of the proficiency cut score on the state test.
2. The proportion of students required to reach the proficiency cut score in a given year, also known as the annual measurable objective (AMO).
3. Whether a confidence interval is applied to proficiency results and its size.
4. The minimum count required for a subgroup to be included in AYP determinations.

Proficiency cut scores and AMOs

A relatively large body of research catalogs differences in state implementations of NCLB and their possible impacts. A number of studies document wide disparities in the state proficiency cut scores (McGlaughlin, Bandiera De Mello, et al. 2008; Peterson and Hess 2008; National Center for Educational Statistics [NCES] 2007; Cronin, et al. 2007; Qian and Braun 2005; Kingsbury et al. 2003;

McGlaughlin and Bandeira de Mello 2002). Others have found differences in the various states' improvement trajectories (Chudowsky and Chudowsky 2008; Porter, Linn, and Trimble 2005; Kim and Sunderman 2004). There is, however, little research available that speaks to the interaction between state proficiency cut scores and these trajectories. For example, some states offset some of the effect of a high proficiency cut score with a back-loaded trajectory of improvement. Other states have lower proficiency cut scores but stricter trajectories for improvement. Whether a particular school makes AYP, then, may be as much a function of the improvement trajectory as the standard's difficulty. Little is known about how these interact in any given state.

Confidence intervals

States have the option to apply a confidence interval to their proficiency scores and the majority of states choose to take advantage of this provision (Fulton 2006). Confidence intervals are ostensibly used to account for sampling error. For example, assume opinion pollsters survey voters in the state of Michigan to estimate their support for a highway bond measure. Obviously the pollsters can't call every voter in Michigan, so they take a sample of 1,000 voters that they hope are representative. They find that 47% of the polled voters support the measure. But they also know that if they repeated the survey with a different sample of voters, the estimate could change. A confidence interval is calculated (based on the number of voters polled) to show how greatly results might vary if the population were resampled. If the poll reports a 95% confidence interval of +/- 3 percentage points, that means that, were the population resampled, the poll would be expected to find between 44% and 50% of voters supporting the bond.

A confidence interval can also be applied to a school's proficiency rate. For example, assume that McKinley Elementary School is required to reach a proficiency rate of 75% in order to reach its AMO and make AYP, but in fact it achieves a proficiency rate of 71%. Assume further, however, that a 95% confidence interval of +/- 6 is calculated by the state and applied to the results. Since McKinley's actual proficiency rate of 71% is within 6 points of the target of 75%, the school would meet this AMO.

Rogosa (2003) argues that the very concept of a confidence interval violates the integrity of a proficiency requirement. In McKinley's case, the school's "real" proficiency rate is as likely to be 65% as it is to be 77%, meaning that the school is far more likely to have failed to reach the proficiency target of 75% than it is to have reached the target. Thus, it would be more reasonable to say that McKinley's status is, at best, *undetermined*. When states use confidence intervals for purposes of NCLB, however, the assumption is that McKinley reached the target.

Other researchers question whether the very concept of the confidence interval is misapplied. Confidence intervals are normally used to compensate for sampling error, but state tests are not administered to a sample of students within a school—they are administered to 95% or more of the eligible students. Thus, the most common justification for the use of confidence intervals wouldn't be appropriate when applied in these circumstances. (M. West, personal communication 2008). This generally leads to an alternate justification for use of the confidence interval, namely, that the state test represents a sample of student performance at a single time, with results possibly varying if students were resampled on a different date. To extend the analogy to opinion polls and voting, this is akin to arguing that election results should be subject to a confidence interval; if the difference in votes between two candidates is within some confidence interval, we should ignore the outcome and revote because the results might be different if we voted the following Tuesday.

The states that employ confidence intervals typically use ranges between 95% and 99% probability, where higher probability means a larger margin around the target value. The differences in the size and application of confidence intervals by the various states can lead to vastly different AYP findings (Erpenbach and Forte 2005; Simpson, Gong and Marion 2005; Porter, Linn, and Trimble 2005). Porter and colleagues found, for example, that the application of a 99% confidence interval increased the proportion of schools that would make AYP in Kentucky schools from 61% to 90% in 2003. The effect of the confidence interval is especially great for small schools or subgroups. In these circumstances, a school with a proficiency rate far below the actual goal may meet the standard if a large confidence interval is employed.

Minimum subgroup sizes

For purposes of NCLB, schools are accountable for the performance of every subgroup of students that exceeds a minimum size established by each state. These requirements vary widely from as few as five students to as many as one hundred or even more. The number of subgroups contained within a school is influenced by three factors: the size of the school itself (a school of 1,000 students with a 10% Native American population is likely to be required to count this subgroup although a school of 100 students with the same proportion of Native Americans will not); the ethnic diversity within the school; and the state's minimum *n* (number of students in sample) requirement. The requirement that proficiency targets be met by all accountable subgroups has led to considerable debate on whether this results in a "diversity penalty" in which racially integrated schools face more difficulties in reaching AYP than more homogenous schools.

Several previous studies (U.S. Department of Education 2006; Kim and Sunderman 2004; Novak and Fuller 2003; Kane and Staiger 2002) have found that schools serving diverse students were at higher risk for failing to make AYP. In a critique of these studies, Rogosa (2005) claimed that the diversity penalty has been overstated, in part because in many low-income schools, different subgroups may have the same membership. In an inner Los Angeles suburb, for example, the Hispanic/Latino, low-income, and limited English proficient (LEP)[1] subgroups may essentially be composed of the same students, meaning that the proficiency outcome for the Hispanic/Latino students is unlikely to differ from that of the other groups.

Moreover, the term "diversity penalty" is itself problematic, because it can imply that holding educators accountable for failing to educate traditionally disadvantaged children is somehow unfair. It is perhaps fairer to question whether accountability and sanctions should be targeted toward poorly performing subgroups as opposed to the entire school (e.g., offering choice to the students in a failing subgroup rather than the entire school).

Still, there are many schools in which the general student population meets its AMO, yet the school fails to make AYP because of the performance of a single subgroup. In 2004, for example, a report from the U.S. Department of Education (2006) found that in 23% of cases schools failed to make AYP because a single subgroup missed an AMO.

The Need for This Study

Ultimately the interactions among the state standards, proficiency trajectory, confidence interval, school enrollment, and minimum subgroup size determine whether a school makes AYP. But, even though it's evident that the standards and rules differ greatly across states, it's extremely difficult to judge or compare the effect that these differences have on the results for individual schools. If a state's application of these rules leads to an overly permissive environment in which nearly all schools, no matter how deficient, make AYP, then we might say that NCLB produces an *illusion* of educational equity. If the application of these rules leads to great inconsistency in the way similar schools are judged across states, it might be more persuasive to argue that these differences lead to unreliable decisions and a subsequent waste of resources. Then again, if AYP findings are fair and consistent *in spite* of differences in applying the rules, we could argue that these complex processes, although messy, still produce the desired result.

Alas, we have found no research to date that examines the interactions between the difficulty of the proficiency standards and the various rules across states. We intend for this study to fill a critical gap in the research by helping policy makers evaluate the consistency of proficiency expectations across states, and determine whether NCLB is consistent in its effect.

[1] Note that we use "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup.

METHODOLOGY

In this section, we give a brief overview of the methods we used to conduct this study. Appendix 1 contains a complete description of our methodology.

Research Question

The purpose of the study was to explore how differences in the various state implementations of NCLB—in this case differences among the states in proficiency cut scores, AMOs, subgroup sizes, and confidence intervals—might interact to affect the AYP status of 36 schools. To address this question, we applied the proficiency cut scores of 28 states and their key AYP rules to a multistate sample of schools.

Sampling and Overall Approach

To begin we created two samples. The first was a sample of states for which we compared cut scores and AYP rules. The second was a sample of schools for which we used achievement data to evaluate the impact of the various state cut scores and rules on their possible AYP status.

In all, we evaluated 28 states in the study. We included a state in the study if sufficient student records from state testing and Northwest Evaluation Association (NWEA) testing were available to permit a robust estimate of the state's proficiency cut scores in both reading and math for grades three through eight.

Our sample of 36 schools was drawn from seven school systems serving 153 schools and located in six states. It was created to reflect the diversity within the American educational system. The sample included schools large and small from both high- and low-income communities. Some of the sample schools served many ethnic groups, others only one or two. Some educated large numbers of students from special populations and some did not. Our sample included traditional public schools, magnet schools, and charter schools. Across the sample, both student achievement and growth varied greatly. We should emphasize that our goal in creating this sample was diversity and not "representativeness." We tried to create a sample that would allow applying proficiency standards and rules to a wide variety of circumstances. Thus we wanted to know if a high performing, non-diverse school, a low performing, diverse school, or a low-performing homogeneous school would make AYP in more states. Creating a "representative" sample of 36 schools, were that even possible, would not have permitted us to engage in this kind of experimentation.

All 36 of these schools participated in both the appropriate state test and NWEA testing during the 2005–2006 academic year. Because NWEA tests are calibrated to the proficiency cut scores of the 28 states included in the study, we had a means to estimate how students in each school would perform relative to the proficiency cut scores in these states. Thus, we could take a school that may have achieved a 70% proficiency rate in Illinois and estimate what its proficiency rate might have been in Wisconsin, Minnesota, New Jersey, or other states. In addition, we could estimate the proficiency rates for various subgroups within each school. Armed with that information, we could assess whether the proficiency rates achieved by the school and its subgroups would have been sufficient to meet the annual proficiency targets required by all 28 states.

We validated that NWEA estimates of a school's proficiency rate within its own state (based on NWEA tests) closely matched the actual results achieved by the school on their own state assessment. If NWEA's estimates of results for a school are a fair reflection of their actual performance on their own state test, they are also likely to produce reasonable estimates of the school's performance on the tests of other states.

Estimating State Test Results

For *The Proficiency Illusion* (Cronin et al. 2007a), researchers aligned the results on NWEA's Measures of Academic Progress (MAPs) with the proficiency cut scores of 26 states. The alignment procedure that was used is outlined in detail in that report, but briefly, alignment was estimated by comparing the performance of a single group of students who participated in both NWEA testing and their respective state's test. The process used, known as "equipercentile equating," is quite straightforward. Assume that 50% of a group of students achieved proficiency on their state's test. If we find the point on the NWEA scale that represents the performance of 50% of the group, that point would represent the score on the NWEA test that is equivalent in difficulty to the proficiency cut score on the state assessment. The accuracy of this process was validated in a pilot study (Cronin et al. 2007b) which found that the equipercentile equating method generally produced projected results that were within three percentage points of the actual state test proficiency rate for the five-state study group.

Since *The Proficiency Illusion* was published in 2007, NWEA has completed estimates for three additional states (and lost one of the original states), now giving us cut score estimates for 28 states. These estimates allowed us to take a student score on the NWEA assessment in one state, and use that score to project whether the student is likely to be proficient in each of the 28 states studied. From there, we were able to project the number of students in each sample school who were likely to be proficient. We could also calculate estimated proficiency rates for each school and its various subgroups.

Note that we were unable to estimate cut scores for eighth grade students in two states, New Jersey and Texas, because of insufficient data. As a result of this limitation, we compared results for the elementary school sample across all 28 states studied, but limited comparisons for the middle school sample to the 26 states in which we had cut score estimates through grade eight.

Estimating a School's AYP Status

Although NCLB requires each state to achieve a target of 100% proficiency for its schools by 2014, each state establishes annual benchmarks for proficiency that increase as schools draw nearer to this deadline. These benchmarks are the AMOs we mentioned earlier. To avoid sanctions, schools must meet the proficiency rate required by the AMO each year.

In addition to setting the AMOs, states also determine minimum subgroup size, and whether and how to apply a confidence interval to a school's proficiency results. For purposes of this analysis, we used the state accountability plans that were in place as of February 2008 (U.S. Department of Education 2008) to document the rules in place at that time. By applying a state's rules to our example schools' data, we were able to project whether a school within the sample would likely achieve several key elements used to determine AYP within that state.

The entire set of rules governing AYP is very complex and it was not possible, based on the data available to us, to estimate the actual status of schools in the sample against all of the AYP rules for the states. As a result, we focused our evaluation on several key AYP rules:

- We evaluated whether the overall performance of students, which we estimated based on spring 2006 results on the NWEA assessment, met the AMOs that the state had set for the 2007–2008 academic year.[2]

- For all ethnic subgroups with counts that exceeded the minimum subgroup size for evaluation, we determined whether their performance, as estimated on the spring 2006 NWEA assessment, was sufficient to meet the proficiency target the state set for the 2007–2008 academic year.

- All students with disabilities (SWDs) were included in the school's sample if they also took some form of their state's assessment. If the count for this subgroup exceeded the minimum subgroup size for evaluation, we determined whether the performance of this group met their AMOs.

[2] As indicated, this report builds on *The Proficiency Illusion* (2007), which used 2005–2006 NWEA data to estimate proficiency cut scores in 26 states. At the time, those were the most recent NWEA data available, and we were unable to update the estimates based on newer data for this report. However, by comparing the 2005–2006 data to the 2007–2008 AYP rules from each state, we're able to use states' most recent annual proficiency targets, which have increased quite dramatically since 2006.

- All students reported as LEP pupils by their schools were included in the school's sample if they also took their state's assessment. Once again they were evaluated against the AMOs if the size of the group exceeded the minimum size.

- All students who were reported by their schools as eligible for free or reduced lunch were included in the sample if they also took their state's assessment. This subgroup was evaluated against the AMO when its count exceeded the minimum size.

- For states that used confidence intervals as part of their AYP calculation, we applied the calculation in circumstances when a subgroup's performance fell short of meeting the required proficiency rate.

To make AYP, elementary and middle schools must also test 95% of their eligible students and meet a standard related to an alternate indicator (generally daily attendance). Data were not available to allow us to evaluate the performance of the sample schools in relation to these two indicators.

Schools that fail to meet an AMO can still make the AYP requirements through a "safe harbor" provision in NCLB. To do this, a school must reduce the number of nonproficient students within a failing subgroup by at least 10% relative to the previous year. We did not evaluate the safe harbor provision as part of this study. As a result, readers should expect that some schools that failed to make AYP in our study might make it in real life.

This methodology allowed us to estimate the proficiency results and status relative to several key AYP rules for each of the 36 schools in the sample. Metaphorically speaking, we were able to drop a school that made AYP in California into states like New Mexico, Illinois, and New Jersey and estimate whether that school would also make AYP there, based on that state's AYP rules.

FINDINGS

How do NCLB's allowances for state discretion affect AYP determinations? To answer this question, we start at the end of the story, by first reporting how our sample of schools performed in the various states relative to making AYP. Next, we explain the components that contributed to this judgment.

How the Sample Performed Relative to State AYP Requirements

Table 1 summarizes the performance of our elementary and middle school samples in making AYP in 2008 across the 28 states we studied. With 18 elementary and 18 middle schools, there were 504 opportunities to make or not make AYP at the elementary level (18 schools x 28 states) and 468 opportunities at the middle school level (18 schools x 26 states).

Table 1. Proportion of schools in the sample that met AYP requirements in 2008

School type	Number and percentage of schools making AYP
Elementary schools	159/504 (32%)
Middle schools	52/468 (11%)

The table shows that our elementary schools made AYP less than one-third of the time. But our middle schools did even worse, making AYP in just over one in ten cases.

Within the elementary school sample, the number of schools that made AYP varied greatly by state. In Massachusetts and Nevada, only one school made AYP, while in Wisconsin, 17 of the 18 schools did (Figure 1). To rephrase, in Massachusetts and Nevada, almost none of

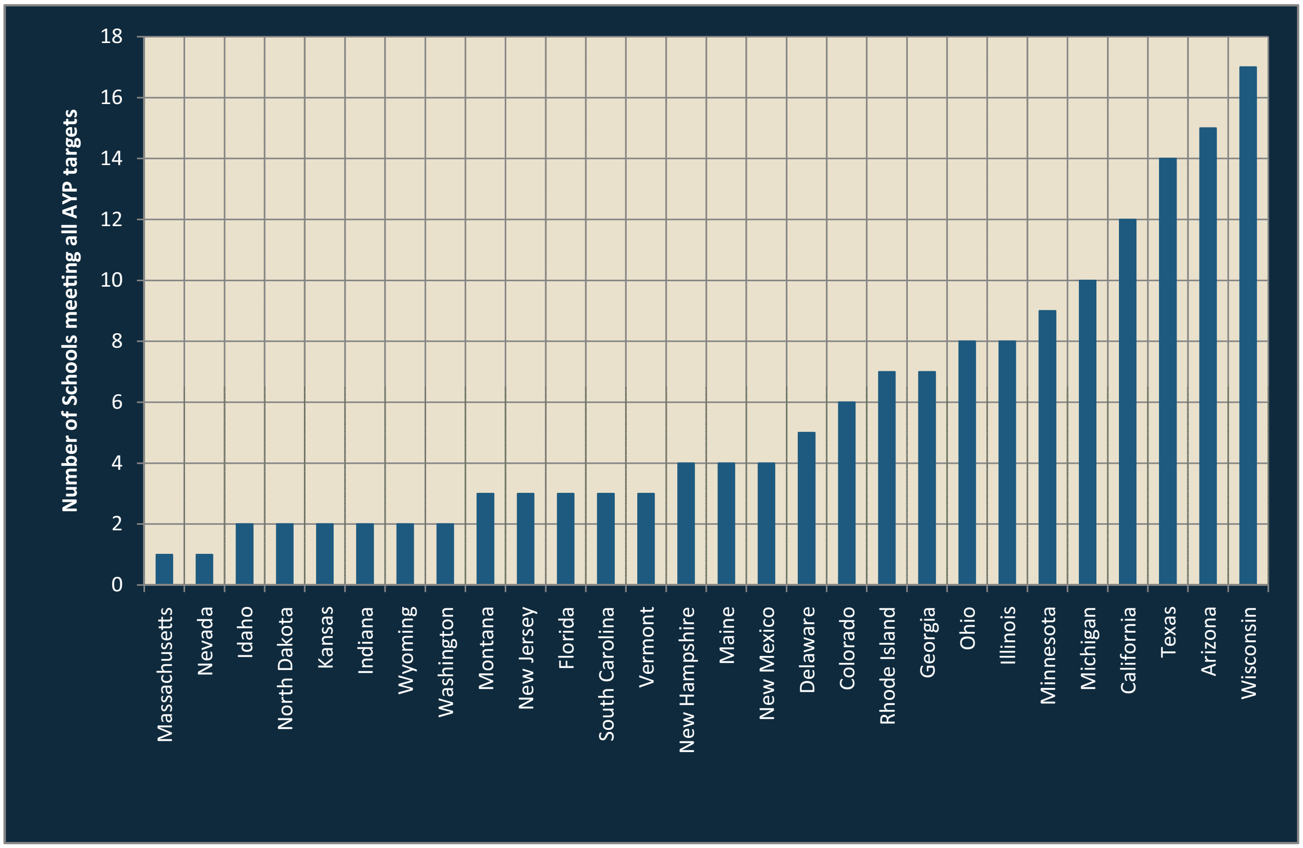

Figure 1. Number of schools in the elementary school sample making AYP by state (2008)

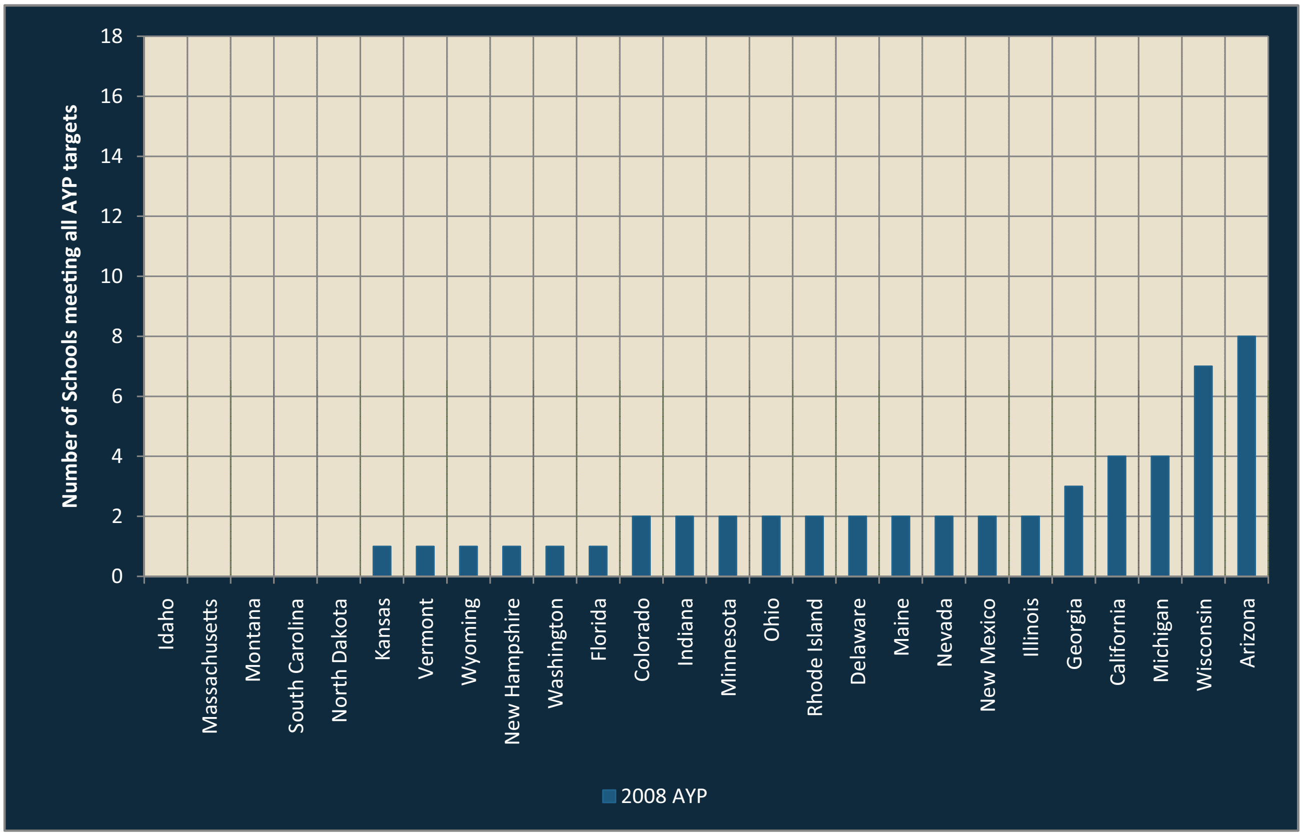

Figure 2. Number of schools in the middle school sample making AYP by state (2008)

Note: Texas and New Jersey are not included in the middle school analysis since cut score estimates for 8th grade were not available in these states.

the elementary schools in our sample made AYP, while in Wisconsin, almost all of them did. **Keep in mind that these are the exact same schools.**

There was more consistency across states with the middle school sample because the vast majority of schools failed to make AYP in most of the states (see Figure 2). In 21 of the 26 states we studied, two or fewer schools met the 2008 AYP requirements. In no state did half of the middle schools meet the 2008 AYP requirements.

The disappointing performance of the schools in the sample led to the questions that ultimately drove the study. For the elementary school sample, why were the AYP outcomes for the group so different across states? For the middle school sample, why did so many fail to make AYP?

The answers to these questions are found in an analysis of three factors that affect whether schools make AYP. These are:

1. The interaction between proficiency cut scores in math and reading and the difficulty of the AMOs;

2. The application of a confidence interval (i.e., margin of error); and

3. The performance of various subgroups, and whether they count for accountability purposes. These subgroups include low-income students, traditionally disadvantaged minorities, limited English proficient (LEP) students, and students with disabilities (SWDs).

In the following subsections, we discuss each of these factors in turn.

The Interactions between Cut Scores and AMO Difficulty (Factor 1, Part 1)

The likelihood that a school will meet an annual target

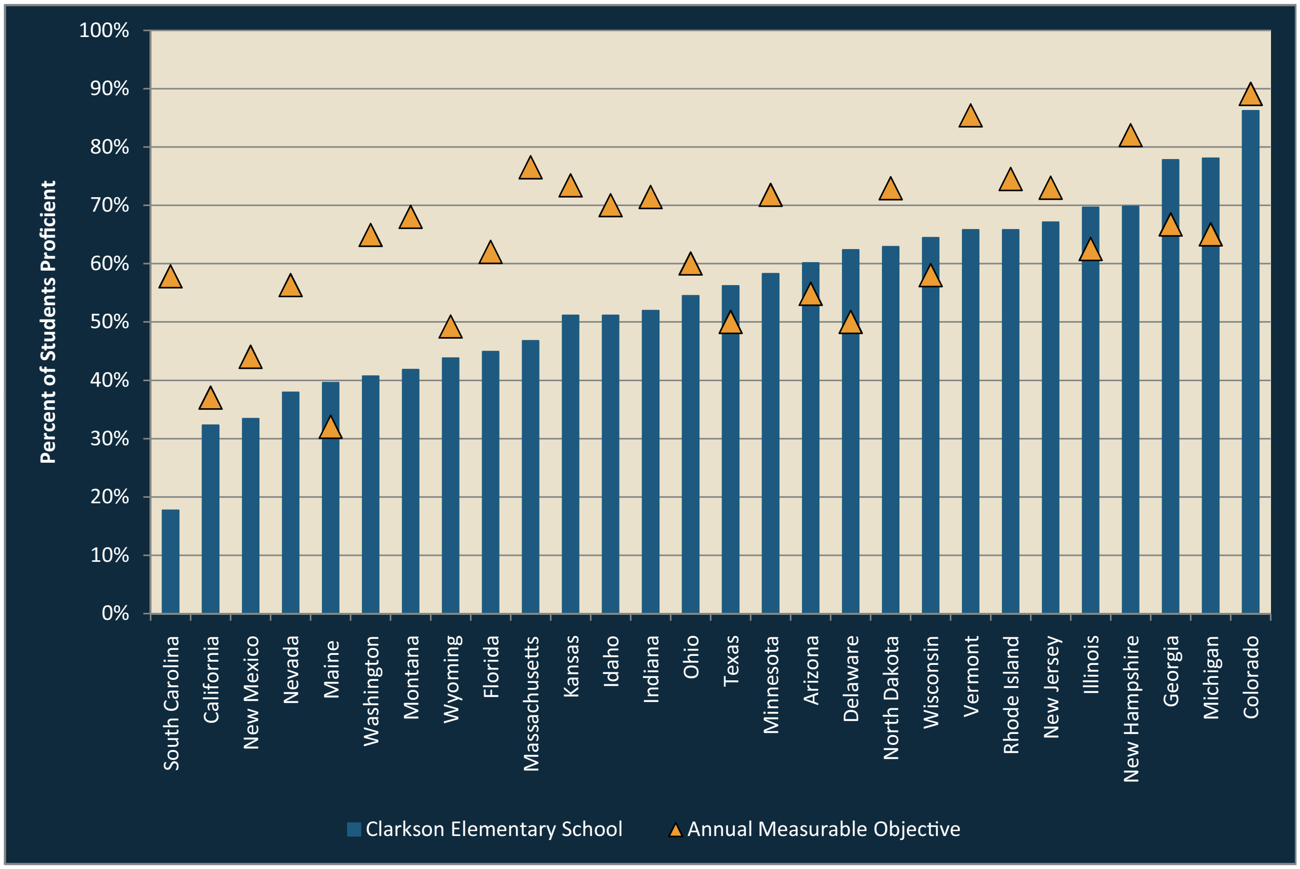

Figure 3. Math proficiency rate of Clarkson students relative to 2008 AMOs

Note: The length of the blue bar represents the percentage of Clarkson students who would be considered proficient in each state. The orange triangle represents the Annual Measurable Objective, or percentage of students required to be proficient in 2008 for the school to make AYP.

is strongly affected by two variables. The first is the difficulty of the test itself. In this case, we aren't talking about the content of the test (which is outside the scope of this study) but instead how difficult or easy it is for students to reach its passing score. The AMOs (i.e., the proportion of students in the school—and in each of the school's subgroups—that must pass the test each year) make up the second variable.

You can have an easy test and a difficult objective. For example, requiring a golfer to make a two-foot putt would be an easy proficiency test in that sport, but asking the same golfer to make 100 two-foot putts in a row would be a difficult objective.

The Case of Clarkson Elementary – Inconsistent proficiency rates and annual targets send conflicting signals

To illustrate this interaction, consider the case of one of our sample schools, Clarkson Elementary, a very diverse school serving primarily low-income students. Ninety-five percent of Clarkson students come from traditionally disadvantaged minority groups (African American, American Indian, and Hispanic/Latino), and 87% qualify for the low-income subgroup. Clarkson is the lowest performing elementary school in the sample. When compared to the NWEA norm group—a sample of over 1.2 million students who attend schools in 32 states (NWEA 2005)—Clarkson students perform, on average, 9.4 scale score points below the norm group's median in math and reading. This would mean that a typical sixth grader at Clarkson performs midway between the fourth grade and fifth grade NWEA norm median in these subjects. In our study, fall to spring scale score growth among Clarkson students was the lowest among the sampled elementary schools; its students attained only 55% of the average growth of students who started with equivalent scores on the NWEA assessments. Setting aside the question of whether Clarkson elementary is a good or a bad school, we would nonetheless expect accountability metrics to identify Clarkson as a school in need of help.

Figure 3 shows the percentage of Clarkson's students who would be projected to reach the proficient level in math (indicated by blue bars) relative to the 2008 AMOs (indicated by the orange triangles) for the states we studied. Clarkson's projected math proficiency rate varied from 18% in South Carolina to 86% in Colorado (which uses "partially proficient" as its standard for NCLB proficiency). Clarkson's proficiency rate was sufficient to exceed the AMOs in 8 of the 28 states studied. So even though this was the lowest performing elementary school in our sample, Clarkson's performance in 2008 would still be considered adequate in eight states. More importantly, we can see very large differences in the percentage of Clarkson students who would be found proficient across states, and equally large differences in how AMOs are set.

In Clarkson's case, the differences in the math proficiency rates and AMOs conspire to send conflicting messages about student achievement based on the state in which the school is placed. If Clarkson were located in South Carolina, for example, its projected results on the state's current assessment (the Palmetto Achievement Challenge Tests, or PACT) would signal that the school's performance is entirely inadequate. Proficiency standards (i.e., the placement of cut scores) in South Carolina are challenging—only 18% of Clarkson students would have passed—and South Carolina's AMO requires 58% of students to pass. The resultant gap (Clarkson's pass rate would need to improve by 40 percentage points just to reach the AMO for 2008) would lead district administrators to conclude that major changes were needed. Overcoming such failure would likely require profound changes in the school's curriculum, culture, and staffing.

When we move Clarkson to Rhode Island, the situation looks far less bleak. Clarkson's math proficiency rate improves from 18% to 67%, a level of performance that fell within a stone's throw of the school's AMO (73%). We can envision incremental improvements to address this kind of gap, perhaps a school improvement plan focused on students' primary deficits. Parents and others reviewing achievement at Clarkson might not believe that performance is that bad, and relatively modest changes might, at least temporarily, fix the school's ailing proficiency rate.

Now, let's move Clarkson to Michigan. Here, math achievement seems to be just fine. More than three-quarters of the students (78%) are projected to achieve proficiency, a level of performance that is well beyond the 2008 AMO (65%). In such a setting, math achievement of the student body as a whole would hardly be a problem, and Clarkson's efforts would be focused on particular subgroups, if any, that may have failed to meet their AMOs.

Unfortunately, things at Clarkson are not fine. Not only is student achievement low, but students are making less progress than their peers. The problem is not limited to small enclaves of minority students, LEP pupils, or students with disabilities either; low achievement persists in all of the school's subgroups. But the messages delivered via accountability systems are highly inconsistent for schools like Clarkson across the country. In some states, the school is on an inevitable path to closure or reconstitution. In others, the problems seem solvable with an educational tweak here or there, and in a few states, there appears to be no problem at all.

Interactions between Cut Scores and AMOs Across the States (Factor 1, Part 2)

As we explained earlier, a school's likelihood of making AYP is affected by the interaction between the proficiency cut scores and the AMOs. Now we examine how this interaction played out in the various states in our study.

Figure 4 illustrates the difficulty of the various state cut scores in math by showing how our sample of eighteen elementary schools performed relative to those targets. In the majority of the states studied, schools are evaluated according to the proportion of students who achieve proficient (or better) on the state test. These states are represented by blue bars in the figure. Six of

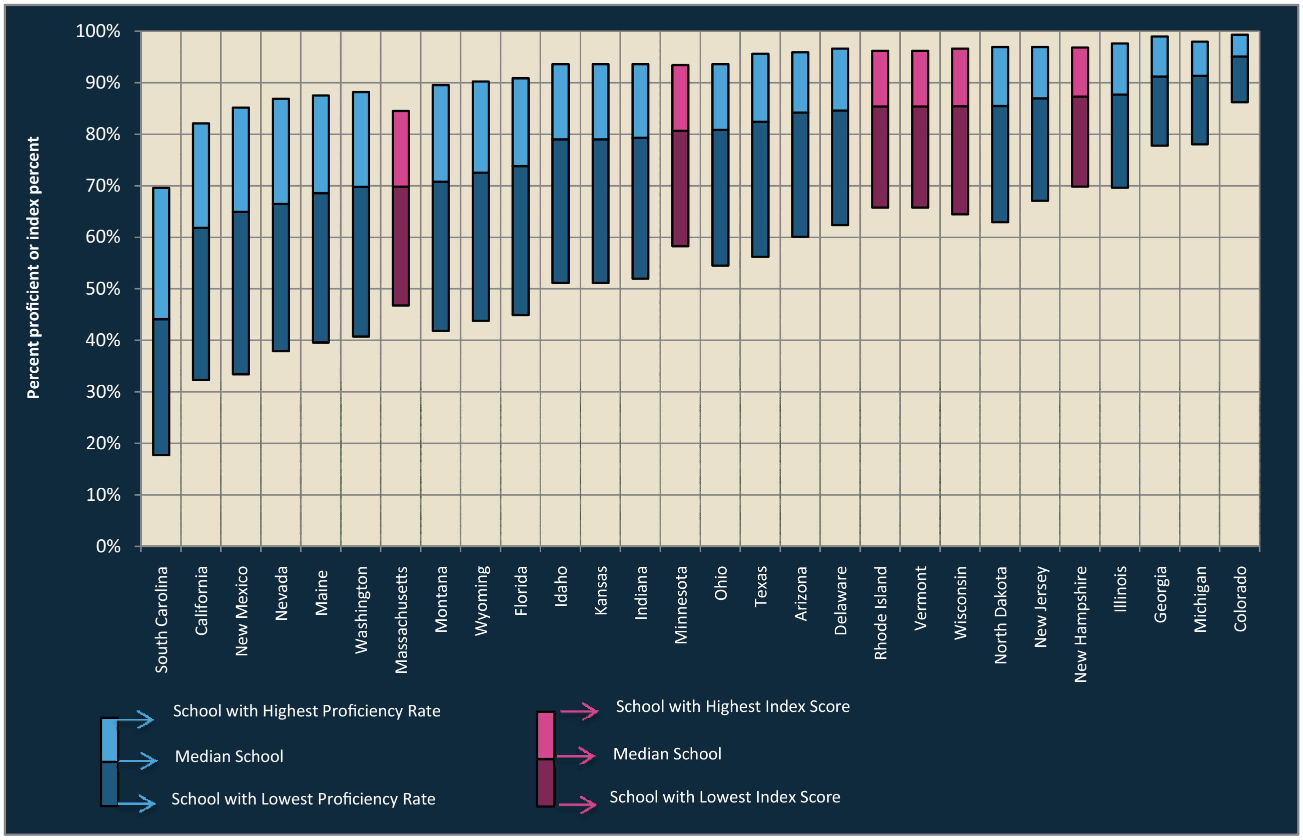

Figure 4. Overall proficiency rates of the elementary school sample in math

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. Magenta colored bars represent states that award students partial credit for achieving at lower proficiency levels.

the states studied (the magenta bars) use an index that gives full credit to students who achieve proficient (or better) and partial credit to students who perform at lower levels. The "index scores" in states using this hybrid model are always higher than the actual proficiency percentage.[1]

The length of the bar in Figure 4 represents the difference in overall performance between the lowest and highest performing sample school in the state. The middle line shows the performance of the median school in the sample. States are ordered by the performance of the median school; consequently, states with higher cut scores are generally located at the left end of the graph, and those with lower cut scores at the right. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). By contrast, in Colorado, the lowest performing school achieved 88% proficiency, the median school achieved 95% proficiency rate, and the highest performing school achieved 99%.

[1] The six states studied that use an index are Rhode Island, Massachusetts, Minnesota, Vermont, Wisconsin, and New Hampshire. The index gives full credit to students who achieve proficient (or better) and partial credit to students performing at lower levels. Consequently, the resultant score in states using this "hybrid" model is always higher than the actual proficiency percentage (giving students partial credit for achieving lower proficiency levels is obviously better than no credit, at least for the schools' ratings). The index provides a fair amount of help when annual targets are below 50%; however, once targets rise above 75%, the index has far less impact.

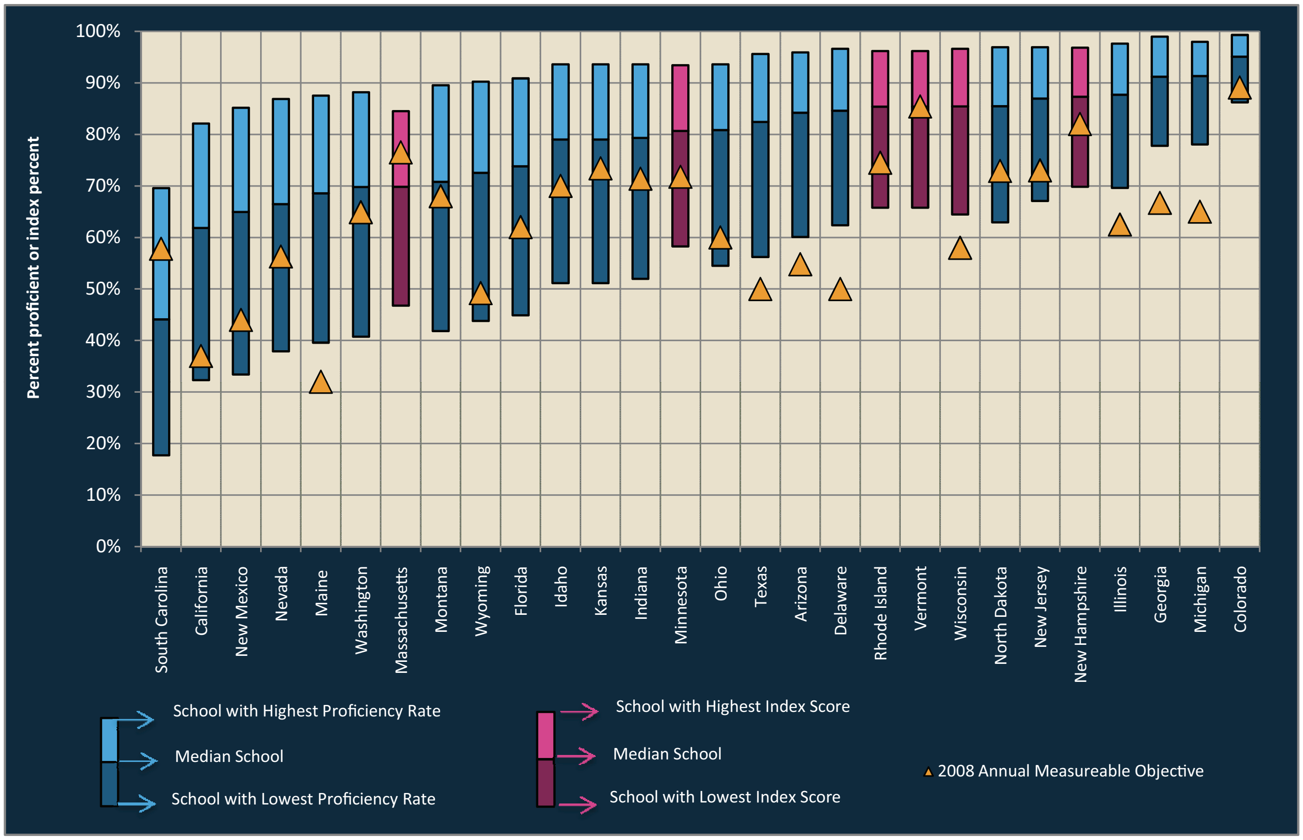

Figure 5. Math proficiency rates of the elementary school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that award students partial credit for achieving at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007-2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

Put another way, fewer than half the schools in our sample would have achieved a 50% proficiency rate if the schools were placed in South Carolina. Had these same schools been located in Georgia, Colorado, or Michigan, the top half of schools would all have achieved estimated proficiency rates greater than 90% (in each of those states, the line dividing the dark and light blue sections of the bar is above 90%).

It's no surprise that the proficiency rates varied from state to state in this study. This finding is consistent with any number of previous studies (McGlaughlin, et al. 2008; Cronin, et al 2007a; National Center for Educational Statistics 2007; Kingsbury, et al. 2003). But the cited studies reflect only one dimension of the assessment, the difficulty of the cut score. The difficulty of the AMOs must also be considered, as we've done in this research.

Figure 5 adds the 2008 AMOs (orange triangles), which show the percentage of students who must be proficient in order for the school to make AYP. The placement of the AMO triangles allows us to see the proportion of the sample that met its target. We can see, for example, that South Carolina's 2008 AMO requires a proficiency rate of 58%. About one-quarter of the sample schools achieved this rate of proficiency. This tells us that South Carolina's proficiency cut score is high relative to the other states and that its AMO is also quite challenging.

Our Michigan results showed the opposite case—Michigan's AMO requires a proficiency rate of 65%, but all

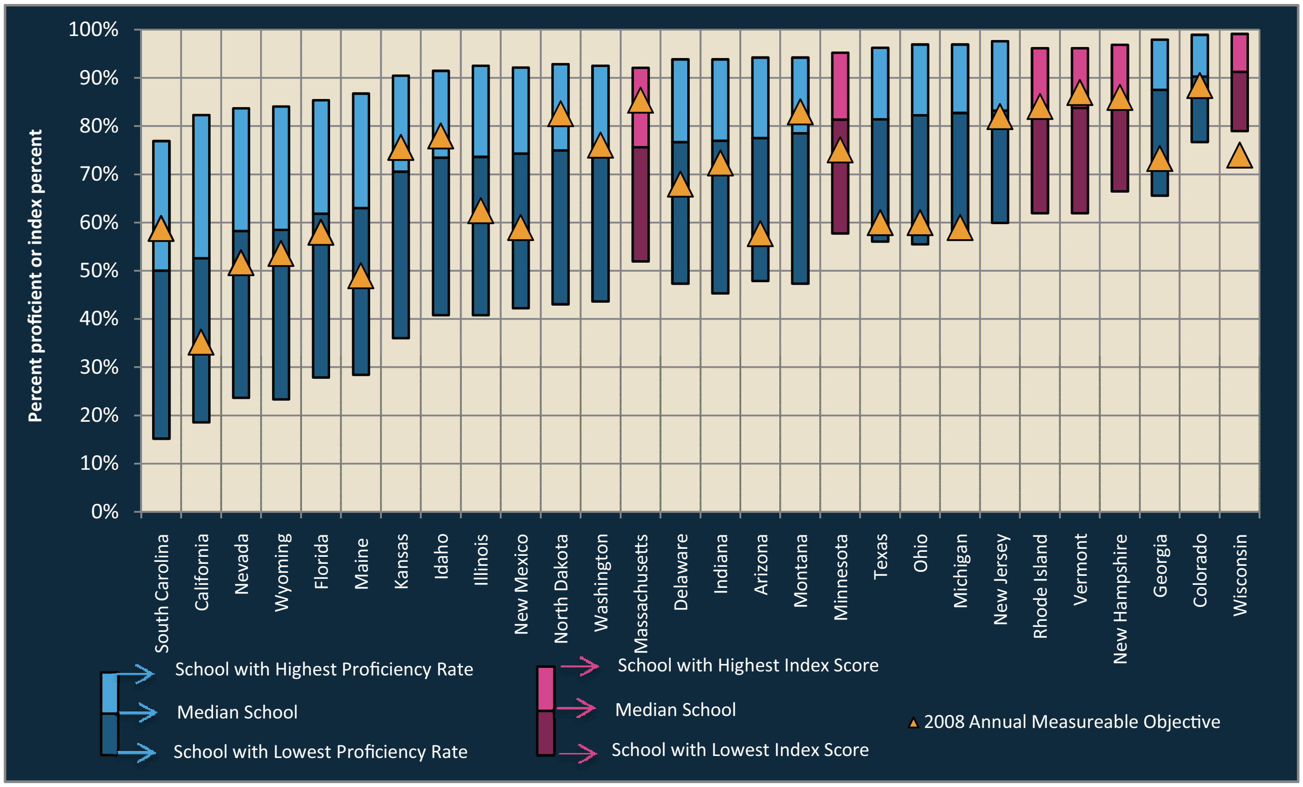

Figure 6. Reading proficiency rates of the elementary school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that give students partial credit for achieving at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007-2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

schools in the sample achieved well beyond this level (indicated by the blue bar floating above the AMO triangle). Keep in mind that we're referring here to schools as a whole reaching their AMOs; we haven't yet considered the impact of subgroup performance. Thus, not only is the Michigan cut score low relative to the other states (remember that states with lower cut scores generally appear on the right), but its AMO is low as well. We could contrast Michigan with Colorado, which reports higher proficiency rates than Michigan (primarily because Colorado gives credit for "partially proficient" students), but has a considerably higher AMO (compare the placement of the orange triangles).

Schools must meet AMOs in both math and reading, so Figure 6 shows the results for the elementary school sample in reading. In general, the AMOs for reading are higher than those for math in the elementary school sample. Although all schools met the math AMOs in eight states (see Figure 5), there was only one state, Wisconsin, in which the entire sample met the reading AMO (indicated by the magenta bar floating above the AMO triangle). In 8 of the 28 states, fewer than half of the schools achieved the AMOs.

Once again, states with relatively low cut scores do not always have easy AMOs. Colorado's AMO was achieved only by about half of the sample, while the AMOs for Wisconsin and Georgia—other states with low cut scores—were achieved by all (Wisconsin) or nearly all (Georgia) schools (note placement of the orange triangles in Figure 6).

Math and reading proficiency rates for the middle school sample were typically lower than those for elementary schools, but AMOs in the states are set at a level that mitigated some of these differences. In seven states (Ari-

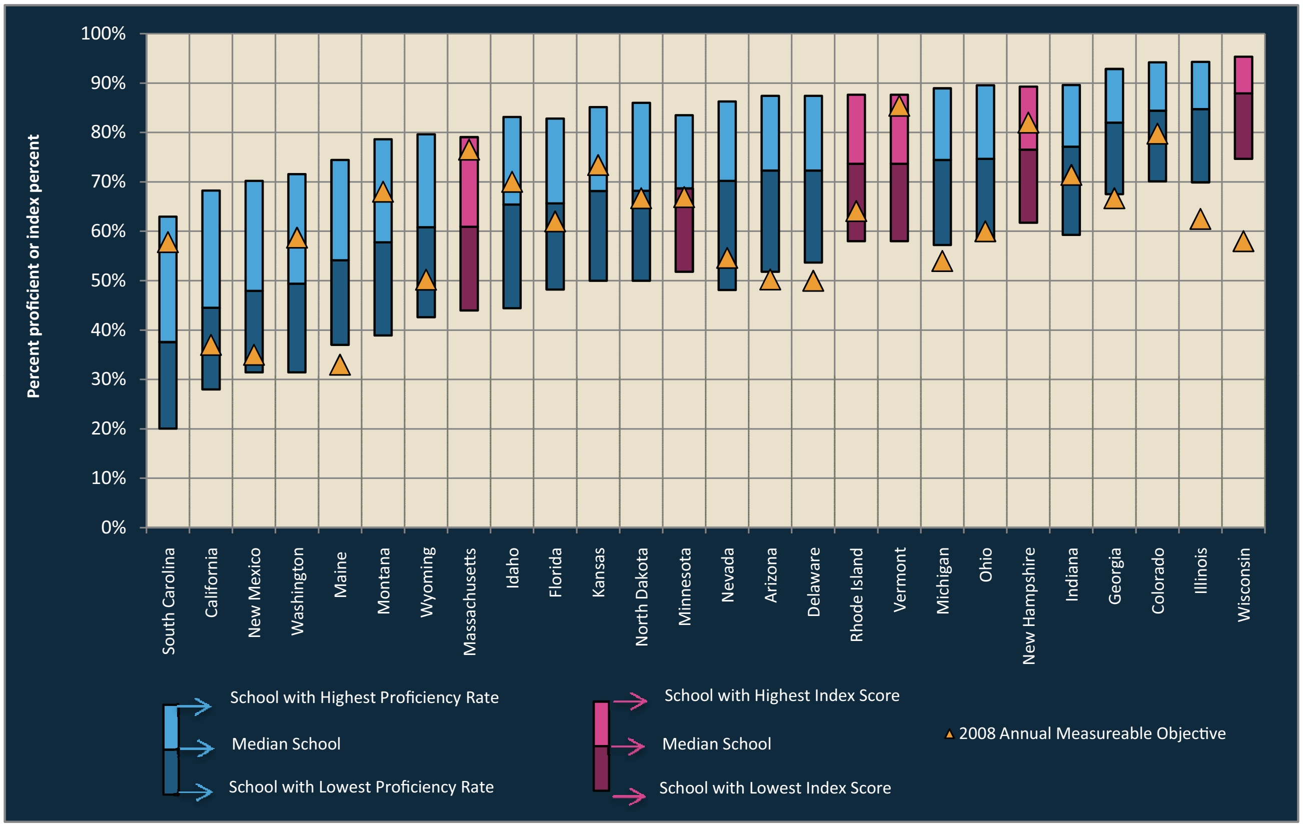

Figure 7. Math proficiency rates of the middle school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that give students partial credit for students who achieve at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007–2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

zona, Delaware, Georgia, Illinois, Maine, Michigan, and Wisconsin), all middle schools met the 2008 math AMOs (Figure 7), and in six states (Arizona, Georgia, Illinois, Michigan, Ohio, and Wisconsin), all middle schools met the reading AMOs (Figure 8). (Again, keep in mind that these results are for schools overall, not for individual subgroups.)

In a few states, however, the AMOs are very challenging. The vast majority of the sample middle schools fail to meet the math AMO in South Carolina (Figure 8). In two of the states (Massachusetts and Vermont) that use hybrid indexes, the majority also failed to meet the math AMOs (note how the AMO triangle appears at the top of each state's bar). The same is true of the reading AMOs in South Carolina, Idaho, North Dakota, Montana, and Vermont. Vermont's case is particularly interesting because it shares a common state test with Rhode Island and New Hampshire. Despite the use of a common test, more of the sample schools failed to meet the AMO in Vermont than in Rhode Island or New Hampshire because Vermont's AMO is higher.

These projections illustrate the importance of considering the AMOs in assessing the impact of NCLB. Much has been made of differences in the proficiency cut scores among the various states, but it's clear that differences in the AMOs have as much impact on the final AYP determination as the differences in cut scores. Some states with high cut scores have not set AMOs that are difficult for most schools to attain. And some states with low proficiency cut scores have AMOs that many schools would not meet. **It is the combination of these two variables that largely determines how easy or difficult it is for schools to make AYP.**

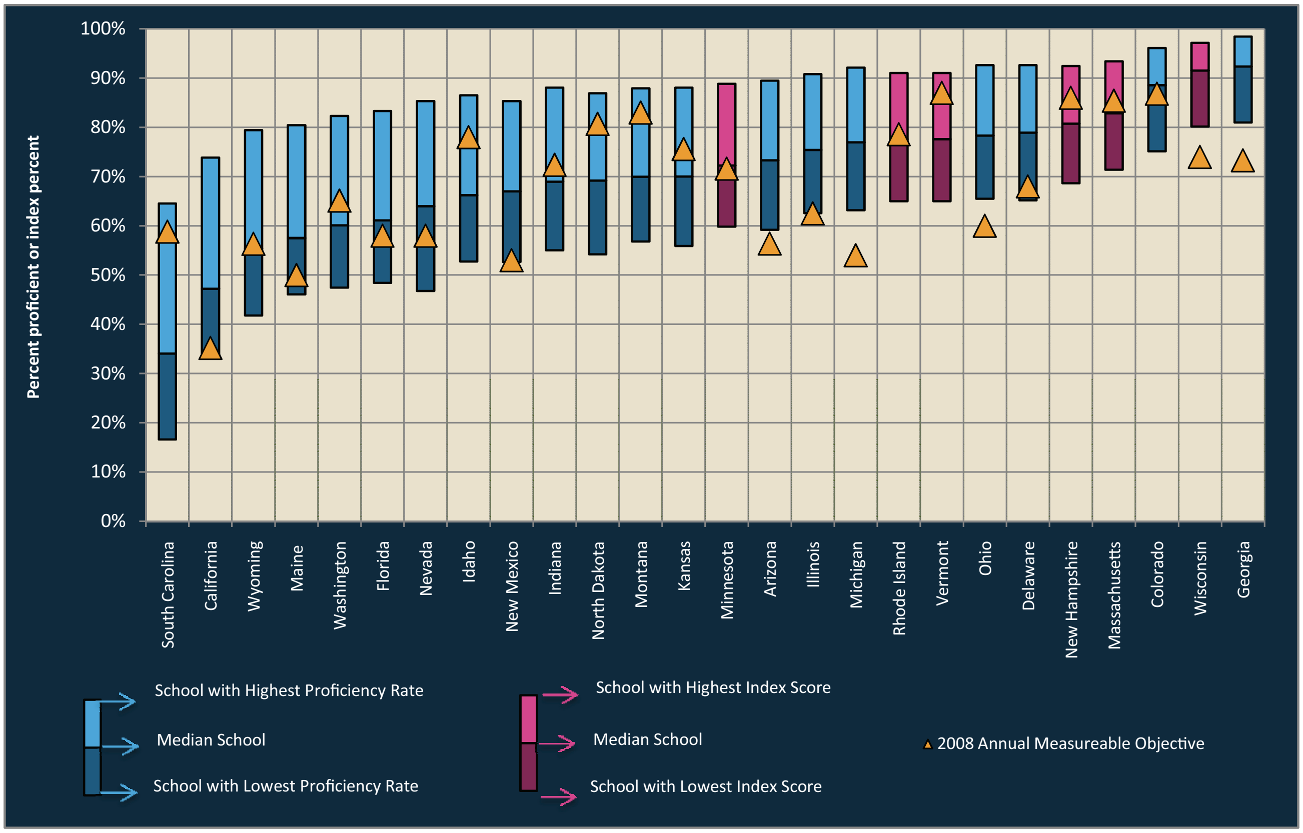

Figure 8. Reading proficiency rates of the middle school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that give students partial credit for students who achieve at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007-2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

The Lowdown on Proficiency Cut Scores and AMOs

The data for Factor 1 lead to several conclusions:

- Disparities in how high or low states set their cut scores lead to large differences in proficiency rates when these various cut scores are applied to a single sample of schools. These inconsistencies make it difficult to know what proficiency really means when comparing states to each other.

- Disparities in the AMOs further cloud interpretation of a school's AYP status. **The combination of big differences in cut scores and AMOs yields a lack of transparency across most state accountability systems.** This murkiness allows a state to correctly claim that its test is more difficult than most, while at the same time permitting nearly all schools, including poor performers, to make AYP because of low AMOs. But other states that have been criticized for their low NCLB proficiency standards (e.g., Colorado), have AMOs that seem reasonable relative to their tests. In these states, many schools may fail to meet their AMOs despite seemingly high proficiency rates.

- In a majority of cases, the math and reading AMOs for the schools' overall populations were met. Despite this, the data will ultimately show that the majority of elementary schools meeting overall proficiency targets ultimately failed to make AYP largely due to subgroup performance; the situation was similar for middle schools. We discuss this further under Factor 3.

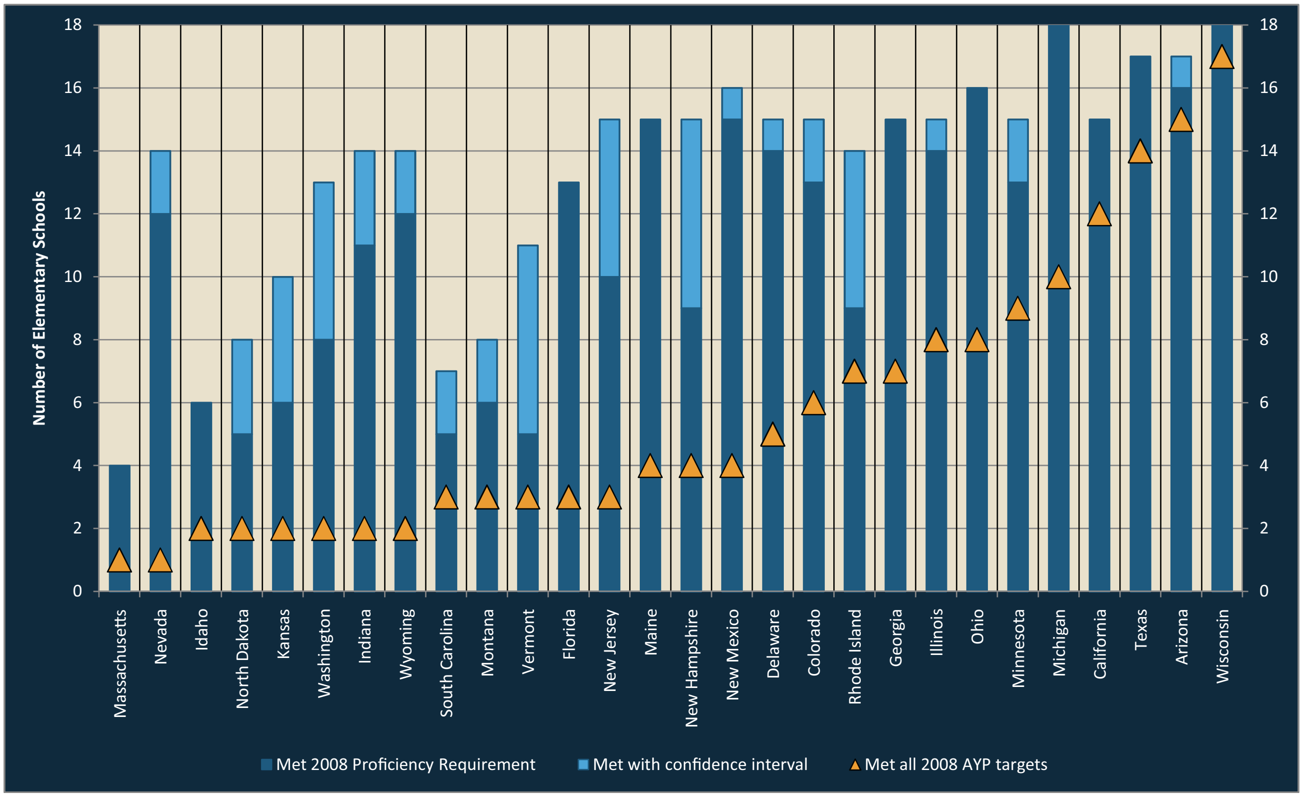

Figure 9. Number of elementary schools meeting 2008 AMOs with and without confidence intervals, by state

Note: The dark blue bars show the number of schools in each state that met their Annual Measured Objectives without employing a confidence interval. The light blue bars show the number of schools that required a confidence interval to meet the target. The orange triangles show the number of schools that ultimately made AYP (with all subgroups meeting their AMOs). For example, the figure shows that despite the fact that 14 elementary schools in Nevada met their math and reading AMOs for their overall student population– two with the help of a confidence interval–ultimately only 1 of those 14 made AYP.

How the Confidence Interval Comes into Play (Factor 2)

Nineteen of the 28 states we studied apply a confidence interval to proficiency test results. For this study, we applied the respective confidence intervals in those states that use them. Table 2 isolates the effect of the confidence intervals and shows how frequently these margins helped elementary schools meet their AMOs for their overall student populations. **In the majority of cases (63%), elementary schools met the AMO without the help of the confidence interval. The confidence interval was required to meet the AMO in about 11 % of cases, and in about 26% of the cases, schools failed to meet the AMO even with the assistance of the confidence interval.**

Table 2. Elementary school sample performance relative to AMOs with and without confidence intervals

Condition	Number of cases and percentage of total
Total measurements (18 schools X 28 states)	504
Cases meeting math and reading AMOs without confidence interval	320 (63%)
Cases meeting AMOs with confidence interval	53 (11%)
Cases not meeting AMOs (even with confidence interval)	131 (26%)

Figure 9 disaggregates the overall proficiency data to show how frequently the confidence interval helped our sample schools meet their 2008 overall proficiency targets in the various states. In 18 states at least one school benefited from the confidence interval in one or both subjects. In five states (New Hampshire, New Jersey, Rhode Island, Washington, and Vermont), five or more schools benefited from it. Overall, however, the vast majority of schools across states that met their AMOs for their overall student population did so without the assistance of a confidence interval.

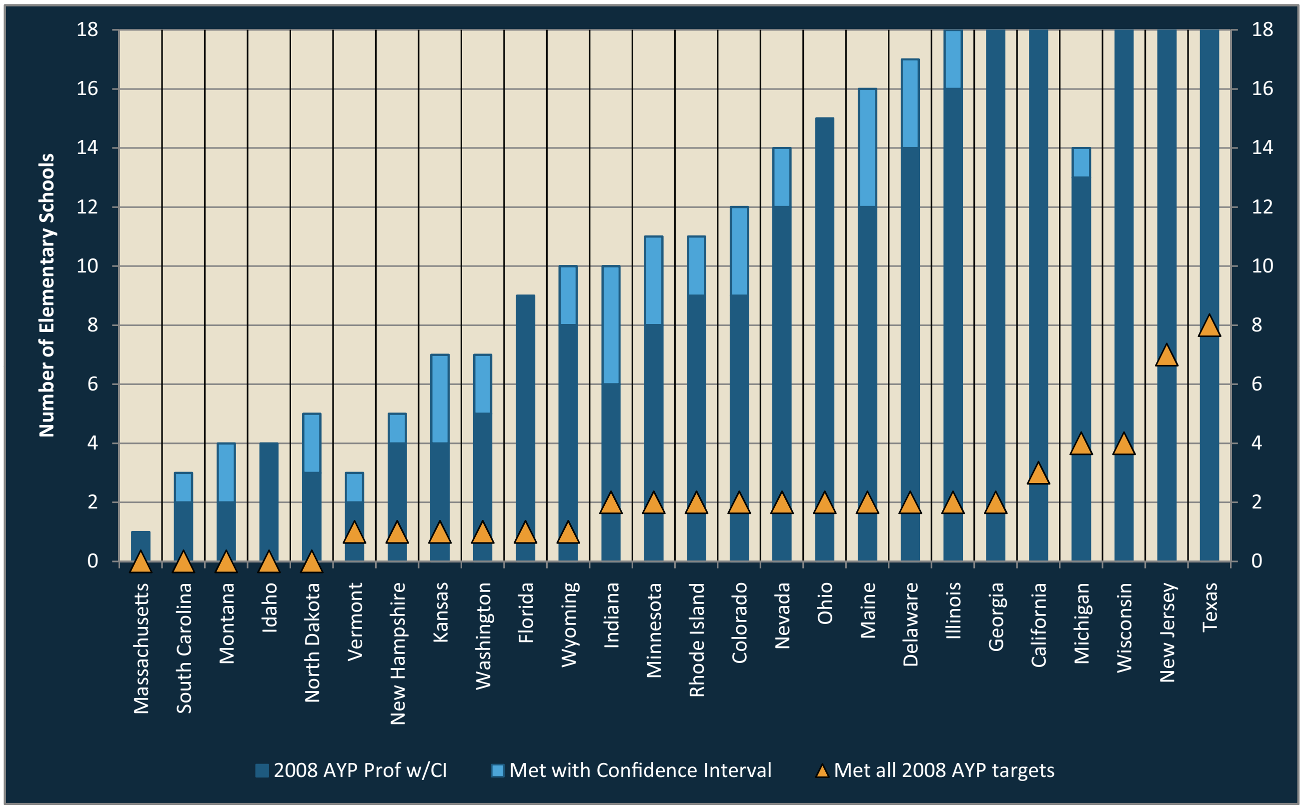

Figure 10. Number of middle schools meeting 2008 AMOs with and without confidence intervals, by state

Note: The dark blue bars show the number of schools in each state that met their Annual Measured Objectives without employing a confidence interval. The light blue bars show the number of schools that required a confidence interval to meet the target. The orange triangles show the number of schools that ultimately made AYP (with all subgroups meeting their AMOs). For example, the figure shows that despite the fact that 14 middle schools in Nevada met their math and reading AMOs for their overall student population–two with the help of a confidence interval–ultimately only 2 of those 14 made AYP.

Table 3 shows that the confidence interval was not quite as helpful to the middle school sample, since it pushed schools past their overall proficiency target in just 8% of cases. In only two states, Indiana and Maine, did the confidence interval help as many as four schools (Figure 10).

Figures 9 and 10 illustrate the effect of the confidence interval when it is applied to the overall population in our sample schools. It is important to remember, however, that when the confidence interval is used, it is not only applied to the overall student population within this study but also to all qualifying subgroups. Thus, the ultimate impact of the confidence interval is larger than the impact depicted in these two figures.

In the analyses appearing in the remainder of this report, confidence intervals were applied to all eligible subgroups in our sample schools, and the results reflect their inclusion. However, we chose not to disaggregate all figures in the report to show the confidence interval's impact because it would have added greatly to the report's length and complexity.

Table 3. Middle school sample performance relative to AMOs with and without confidence intervals

Condition	Number of cases and percentage of total
Total measurements (18 schools X 26 states*)	468
Cases meeting math and reading AMOs without confidence interval	248 (53%)
Cases meeting AMOs with confidence interval	38 (8%)
Cases not meeting AMOs (even with confidence interval)	182 (39%)

*Note: Texas and New Jersey state analyses were not conducted for the middle school sample because proficiency cut score estimates for all middle school grades were not available in these states.

The Lowdown on Confidence Intervals

To summarize our discussion of Factor 2:

- In the majority of cases, schools were able to meet AMOs for overall proficiency without the assistance of a confidence interval.
- In eight to eleven percent of cases, however, the confidence interval allowed schools to meet the AMO for their overall student population.
- When subgroups are considered, the impact of the confidence interval on ultimate AYP determinations is larger.

How the Performance of Student Subgroups Affects a School's Chances of Making AYP (Factor 3)

In this section, we discuss the impact of subgroup performance in general on AYP, including two case studies that show how the state in which a school is located impacts a school's chances of making AYP. Then we turn to a discussion of the performance of specific subgroups, namely low-income students, minority populations, LEP students, and SWDs.

Even if a school's overall proficiency rate is sufficient to meet the AMOs for math and reading, the school must also meet these same targets for each qualifying subgroup to ultimately make AYP. One consistent aspect of NCLB is that within a state, all subgroups must meet the same target. But the minimum size that qualifies a subgroup for separate evaluation differs across states. Some states require groups as small as five students to be evaluated; other states set subgroup minimums at 100 or more (see the State Reports section of this report for the particular requirements of each state).

As shown earlier, it's the combination of cut scores and AMOs that largely determines how easy or difficult it is for schools to make AYP. But a third factor, the minimum subgroup size, is also critical. **As the number of qualifying subgroups within a school increases, each new subgroup introduces another AMO that must be met**. The nature of the qualifying subgroup also makes a difference. It may be easier for a school to address poor performance in an ethnic subgroup than it is to address poor performance among SWDs, or LEP students.

The Case of Chaucer Middle School – A high performing, high growth school runs aground

Chaucer is the highest performing middle school in our sample. Table 4 summarizes the ranking of its students relative to the other middle schools in the sample. Chaucer ranks either first or second in achievement among each of the subgroups in the sample that were large enough for evaluation.

Table 4. Ranking of Chaucer middle school students relative to entire middle school sample

	Student Count	Ranking among middle school sample (reading)*	Ranking among middle school sample (math)*
All students	1118	1st	1st
Low-income students	112	1st	1st
Hispanic/Latino students	135	1st	1st
African American students	31	2nd	1st
Asian students	153	1st	2nd
LEP students	61	1st	2nd
SWDs	88	2nd	1st

* Minimum *n* of 10 students required for consideration. There are 18 middle schools in the sample.

LEP=limited English proficient; SWDs=students with disabilities

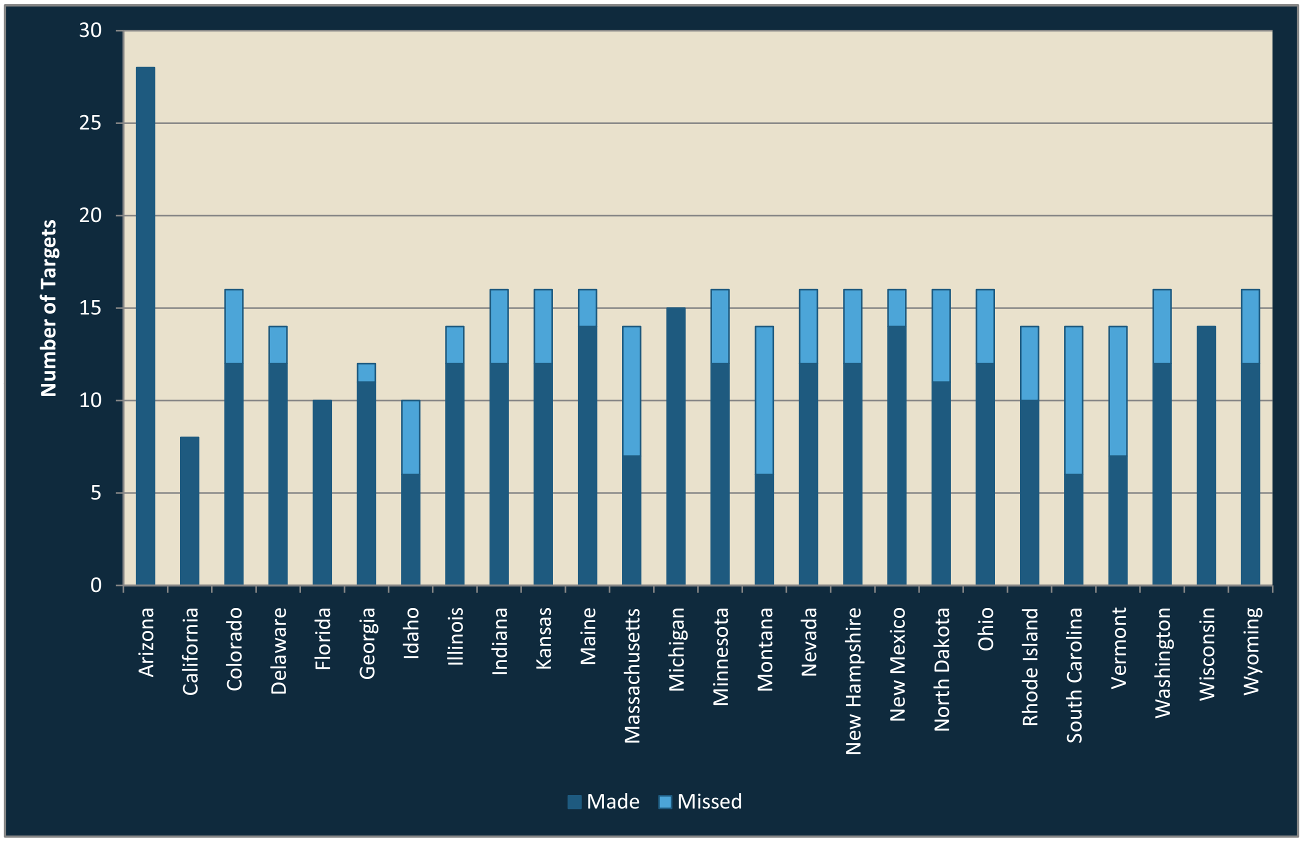

Figure 11. Number of subgroup targets met by Chaucer middle school in 2008

So how did Chaucer perform relative to the states' AYP requirements? Miserably. Chaucer made AYP in only 5 (Arizona, California, Florida, Michigan, and Wisconsin) of the 26[2] states evaluated (Figure 11). What caused this? Certainly not Chaucer's overall performance, which exceeded the annual targets in every state. Was it because of the performance of Chaucer's low-income or minority students? This is a partial explanation. Indeed, Chaucer's low-income subgroup failed to make AYP in six states and one or more of its minority subgroups failed in five states (not shown). This happened despite the fact that all of these subgroups showed above average performance relative to students in the NWEA norm group in their respective grades.

But the biggest explanation for Chaucer's failure is the performance of its LEP students and its SWDs (not shown). The LEP subgroup met its AMOs in only 2 states, failing in 20. (In the other four states, the size of this subgroup fell below the states' minimum for inclusion.) Similarly, the SWDs subgroup made its AMOs in only 2 of 26 states, failing in 21. The irony here is that Chaucer's LEP and SWD subgroups performed better than almost every other subgroup in the sample. So here is a school that is taking students with known learning challenges, presumably providing more effective help to these students than the other schools in the sample, and still failing to make AYP in more than 75% of the cases we studied. In fact, no school in the sample served students in these subgroups better. Chaucer himself aptly described the predicament of his namesake school; "…If gold rusts, what shall iron do?" If a school like *this one* is labeled a failure under NCLB, just where does one think its students should go to be better served?

In short, Chaucer ran aground primarily for two reasons. First, it's at a huge disadvantage because it's judged on

[2] While 28 states are included in the study for elementary school results, we lacked sufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to 26 states.

Table 5. Ranking of Pogesto middle school students relative to entire middle school sample

	Student count	Performance rank among middle school sample* (reading)	Ranking for student growth among middle school sample* (math)
All students	54	14th	4th
Low-income students	26	3rd	5th
White students	41	18th	5th
Hispanic/Latino students	12	7th	4th

* Minimum *n* size of 10 students required for consideration. There are 18 total middle schools in the sample.

whether two subgroups with documented learning challenges—limited English proficient students and students with disabilities— met a fixed and somewhat arbitrary proficiency target, rather than whether it produced strong results and improvement in the performance of these groups. Second, it is a large school in a diverse community, which means that there are many subgroups of students and many of these groups are larger than the minimum *n* size required for evaluation. Large, diverse schools are accountable for the proficiency rate of a large number of subgroups—meaning they have many more targets to meet. On the other hand, smaller schools may be less effective, yet meet AYP because they have fewer qualifying subgroups and fewer targets to hit. Our next example illustrates this problem.

The Case of Pogesto Middle School - Small size benefits a low-performing school

Pogesto, an alternative school serving middle school students, was one of the lowest performing schools in the sample. It ranked 14th out of 18 schools in overall performance in reading and 18th in terms of white subgroup performance in reading (Table 5). Its students averaged about 3.9 scale score points below NWEA's norms, the equivalent of roughly one-half grade level. All Pogesto subgroups with counts greater than ten performed below NWEA norms. On the other hand, growth rates in math at Pogesto were above average; it performed in the top-third of the middle school sample in this regard.

Based on the results for Chaucer, we would expect Pogesto to fail to make AYP in almost every state. But Pogesto made AYP in 15 of the 26 states studied (Figure 12); only one school in the middle school sample performed better. How did this happen?

The answer is simple. With 54 students, Pogesto had fewer students than any of the other middle schools in the sample. Its subgroups are so small that one is rarely large enough to be included. In 19 of the 26 states in our study,[3] we evaluated Pogesto solely on the reading and math performance of its general student body and, in some of these states, on the performance of its white student subgroup. In only seven states (these are the states with more than four subgroup targets in Figure 12) was Pogesto required to meet AMOs with additional subgroups, and in five of these seven states, it made AYP (Arizona, Maine, Minnesota, Nevada, New Mexico).

Pogesto is not a bad school. It is actually an alternative school that serves students who have not performed well

Table 6. AYP designations for Pogesto and Chaucer middle Schools in 26 states

Both made AYP	Pogesto made AYP – Chaucer did not	Chaucer made AYP – Pogesto did not	Both failed to make AYP
4 states	11 states	1 state	10 states

[3] Recall that two states (Texas and New Jersey) were not included in the middle school analysis because of insufficient data.

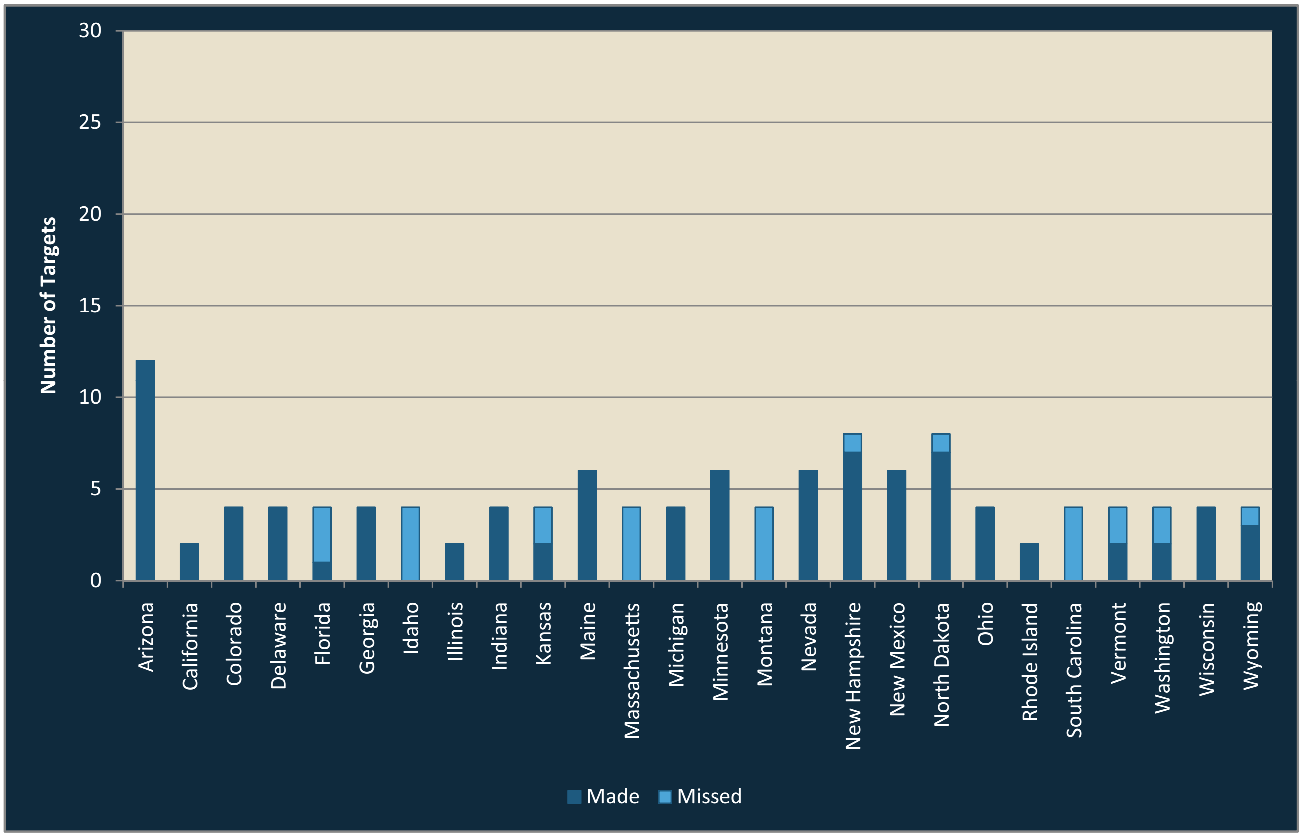

Figure 12. Number of subgroup targets met by Pogesto middle school (2008)

in other settings. Its low-income students performed near the top of the sample (though below the NWEA average) and the school's growth was within the upper third of the schools sampled. Whether Pogesto is a good or bad school, however, is not the point. Instead, the question is whether Pogesto—and other schools in the sample—are judged consistently. The answer is no. In this study, Pogesto was less effective than Chaucer by almost any measure, yet most state accountability systems have indicated otherwise. Indeed, it is remarkable that only one state (Florida) appropriately "passed" the higher performing, higher growth Chaucer while "failing" the lower performing, lower growth Pogesto (Table 6). Even more remarkable is the fact that Pogesto met AYP in 11 states where Chaucer failed to do so.

Again, Pogesto made AYP in most states because it's small and has few subgroup targets to hit, and Chaucer failed because it's large and has many subgroup targets to hit. Next, we isolate the effect of particular subgroups on the study sample.

Performance of low-income students

Even if the overall proficiency rate within a school is sufficient to meet the AMOs for math and reading, schools must still meet these same objectives for each qualifying subgroup in order to make AYP. After white students, the largest of the subgroups is typically low-income students. Table 7 summarizes the performance of this subgroup of students in the elementary school sample.

Table 7. Elementary school sample performance relative to the AMOs for low-income students

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which low-income group was below the minimum subgroup size	55 (11%)
Number of cases in which low-income group met all AMOs	223 (44%)
Number of cases in which low-income group failed to meet one or more AMOs	226 (45%)

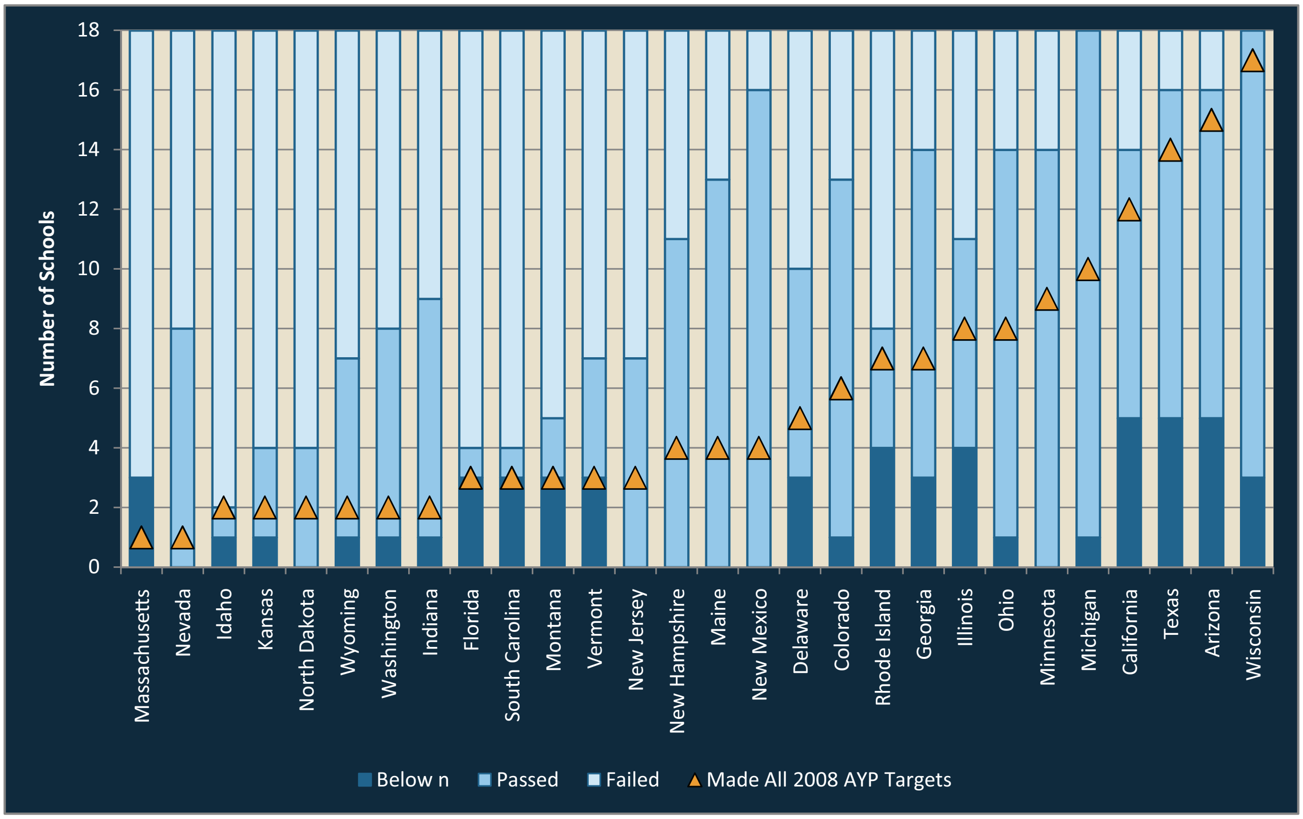

Figure 13. Number of elementary schools meeting 2008 AMOs in math and reading for their low-income student subgroup

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Massachusetts, every elementary school with a qualifying low-income subgroup failed to meet its AMOs. In Michigan, however, every school with a qualifying low-income subgroup passed its AMO. Note, however, that even though all the low-income subgroups met their AMOs in Michigan, only 10 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining eight failed to make AYP because of some other subgroup.

Subgroup counts were below the minimum size in only 11% of our cases. In 44% of cases, the low-income subgroup met all AMOs; it failed one or more AMO in slightly more cases (45%).

Table 8. Middle school sample performance relative to the AMOs for their low-income students

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which low-income group was below minimum subgroup size	27 (6%)
Number of cases in which low-income group met all AMOs	149 (32%)
Number of cases in which low-income group failed to meet one or more AMOs	292 (62%)

Note: While 28 states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to 26 states.

Figure 13 shows how the sample elementary schools fared by state. In one state, Massachusetts, all schools with a low-income qualifying population failed to reach their AMOs (failures are indicated by the light blue bar). In two states, Wisconsin and Michigan, we have the opposite situation; all the sample schools with a qualifying count for low-income students passed their AMOs (indicated by the median shade of blue).

Because the middle schools in our sample are considerably larger than most of the elementary schools, there were only 6% of cases in which the low-income subgroup fell below the minimum *n* size required for evaluation (Table 8). In 32% of the total cases, the school met its required AMO for the low-income subgroup, but schools failed in well over one-half (62%) of the cases.

In four states (Idaho, Massachusetts, Montana, and South Carolina), no middle school with a qualifying

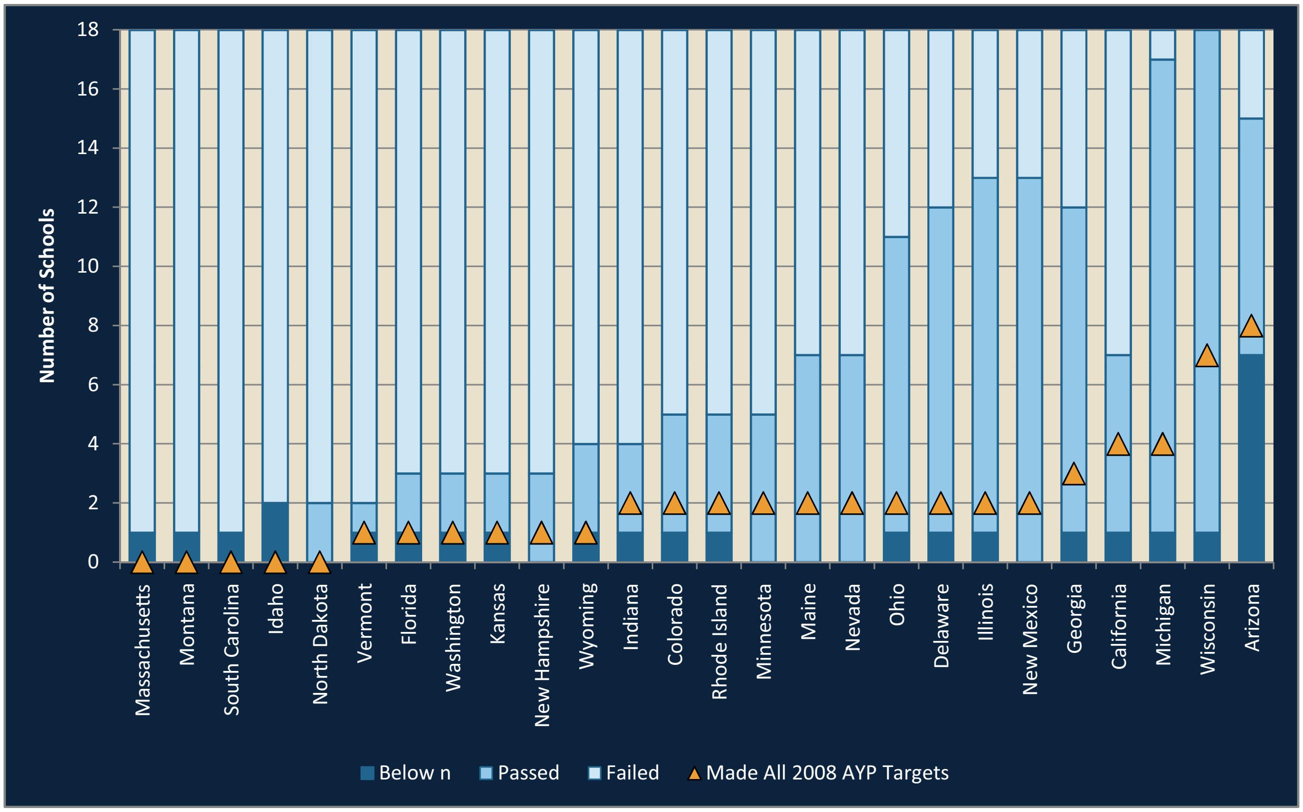

Figure 14. Number of middle schools meeting 2008 AMOs in math and reading for their low-income student subgroup

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Massachusetts, every middle school with a qualifying low-income subgroup failed to meet its AMOs. In Wisconsin, however, every school with a qualifying low-income subgroup passed its AMO. Note, however, that even though all the low-income subgroups met their AMOs in Wisconsin, only 7 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining 11 failed to make AYP because of some other subgroup.

low-income population met the AMOs for that group (Figure 14). There was one state, Wisconsin, in which all sample middle schools with a low-income qualifying population passed. In 18 states, half or more of the low-income subgroups within the middle school sample failed this AMO (note all of the long light blue bars in Figure 14). The AYP performance of the schools provides an interesting contrast. They show, for example, that even in states where the low-income students made their AMO, it did not necessarily help assure a positive final outcome for the school. For example, 13 schools in New Mexico met the AMO for low-income students, and 11 of the 13 still failed to make AYP.

Overall, elementary schools failed to meet the annual targets for the low-income subgroup in 45% of cases, while middle schools failed to meet it in 62% of cases. These failures were not evenly spread across states, but concentrated among about two-thirds of the sample states.

Performance of minority students

Table 9 reports the performance of minority students within the sample elementary schools relative to their 2008 AMOs for reading and math across all states studied. In about 27% of the total cases, schools in the sample had no minority group large enough to meet the

Table 9. Elementary school sample performance relative to the AMOs for their minority students

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which all minority groups were below minimum subgroup size	134 (27%)
Number of cases in which all minority groups met all AMOs	139 (28%)
Number of cases in which some minority groups failed to meet one or more AMOs	231 (46%)

Note: Percentages may not add to 100 due to rounding.

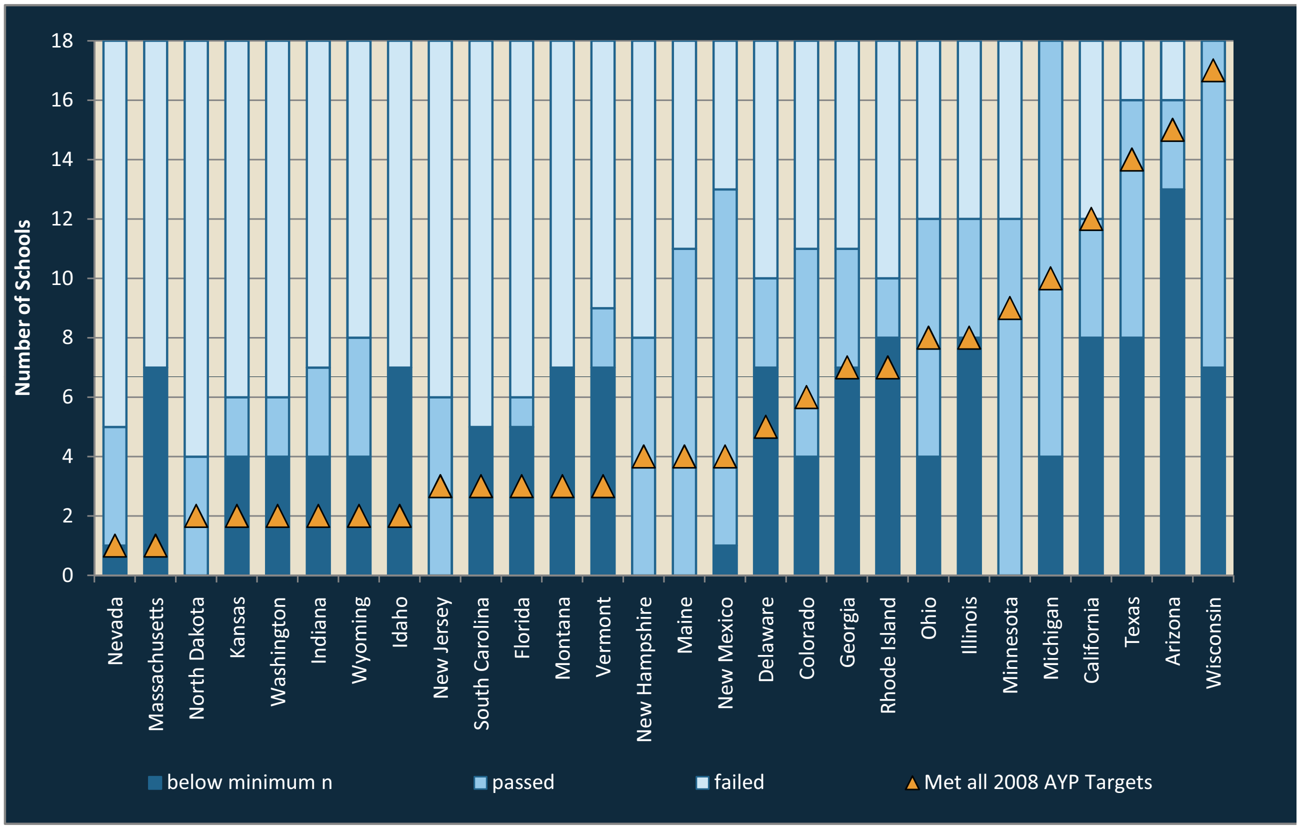

Figure 15. Number of elementary schools in which minority students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Massachusetts, every school with a qualifying minority subgroup failed to meet its AMO. In Michigan, however, every school with a qualifying minority subgroup passed its AMO. Note, however, that even though all the minority subgroups met their AMOs in Michigan, only 10 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining 8 failed to make AYP because of some other subgroup.

minimum reporting requirement. Among the remainder, all qualifying minority groups met their objectives in math and reading in 28% of cases, but in 46% of cases, one or more minority groups failed to meet the objectives in one or both subjects.

Figure 15 shows the distribution of results for the elementary school sample by state. Because of a low minimum *n* size requirement, there were five states in the sample (Maine, Minnesota, New Hampshire, New Jersey, and North Dakota) in which all schools had at least one minority subgroup that exceeded the minimum subgroup size.

There were four states (Idaho, Massachusetts, Montana, and South Carolina) in which all schools with a minority subgroup that met the minimum *n* size failed one or more AMOs. All four of these states had relatively high cut scores. In 13 other states, more than half the schools had at least one minority group that failed to meet an annual target; these states also had cut scores that fell in the upper half in difficulty. But there were also two states, Michigan and Wisconsin, in which all schools

Table 10. Middle school sample performance relative to the AMOs for minority students

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which all minority groups were below minimum subgroup size	40 (9%)
Number of cases in which all minority groups met AMO	103 (22%)
Number of cases in which some minority groups failed to meet one or more AMOs	325 (69%)

Note: While twenty-eight states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to twenty-six states.

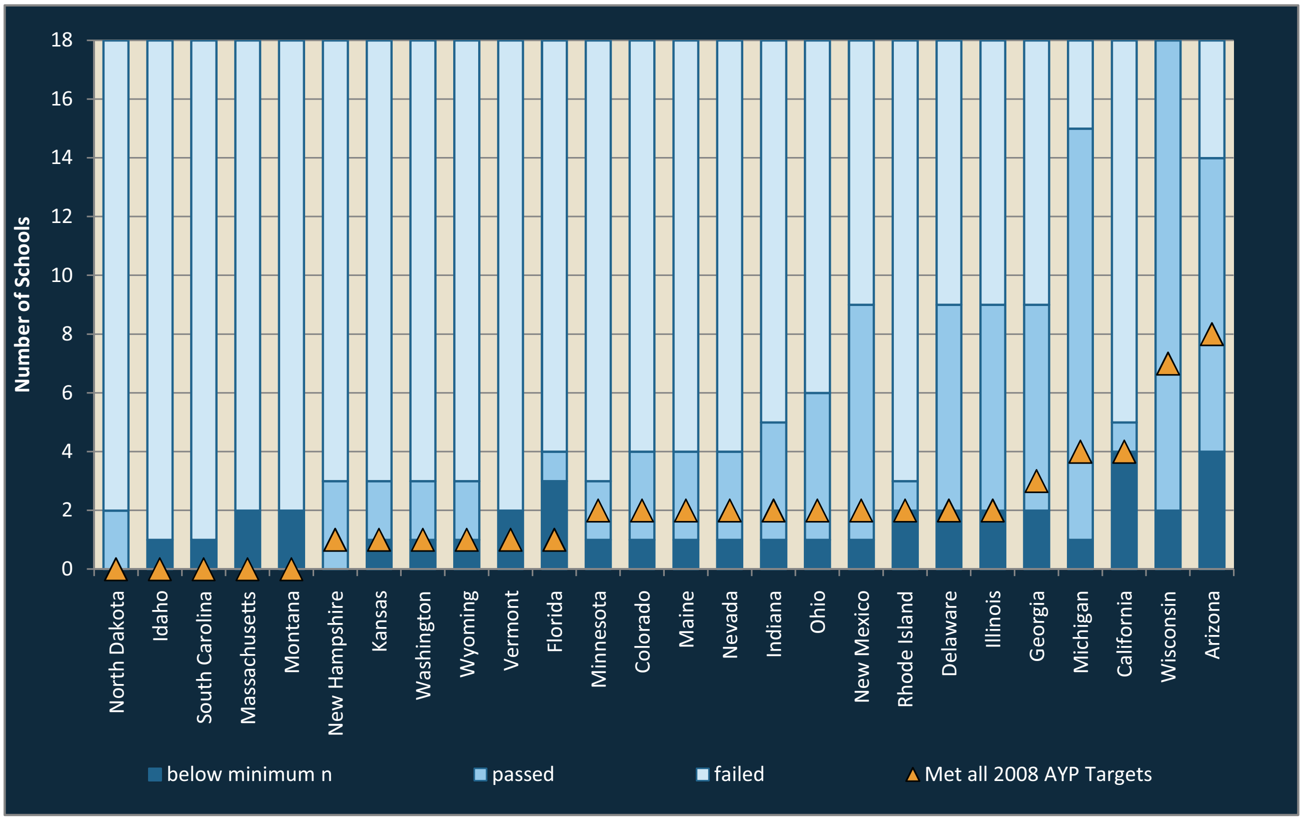

Figure 16. Number of middle schools in which minority students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in North Dakota every school with a qualifying minority subgroup failed to meet its AMO. In Wisconsin however, every school with a qualifying minority subgroup passed its AMO. Note, however, that even though all the minority subgroups met their AMOs in Wisconsin, only 7 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining 11 failed to make AYP because of some other subgroup.

with a qualifying minority group passed. These two states have both lower than average cut scores and lower than average AMOs. Finally, there are several states in which many schools that met the AMOs for their minority students ultimately failed to make AYP on some other basis. In Maine, for example, there were 11 schools in which all minority subgroups met the AMO, yet only 4 of these schools ultimately made AYP. While all schools in Michigan with a qualifying minority subgroup saw those subgroups meet the AMO, 8 of the schools failed to make AYP because of some other subgroup.

Once again, the middle schools in the sample performed worse than the elementary schools. Because middle schools are generally larger than elementary schools, in just 9% of the cases were there no minority groups in a school large enough to qualify as a subgroup—less than half what was found in the elementary school group. Minority groups passed all of their proficiency objectives in 22% of cases, but failed in 69% of cases, a failure rate 22 percentage points higher than the elementary school failure rate (Table 10).

In five of the states (Idaho, Massachusetts, Montana, South Carolina, and Vermont), all middle schools with a qualifying minority group failed to meet that group's targets (Figure 16). In 19 of the 26 states, more than half the middle schools in the sample failed to meet their targets for one or more of their minority groups. The only state in which all schools with a qualifying minority group passed was Wisconsin, but more than half of the schools also passed the targets in Michigan and Arizona. Once again, there are several states in which the minority subgroups of many schools met their AMO, yet the vast majority of schools still ultimately failed to make AYP. In Michigan, for example, all minority subgroups passed in fifteen schools, but only four of these schools ultimately made AYP (indicated by the orange triangle). In Wis-

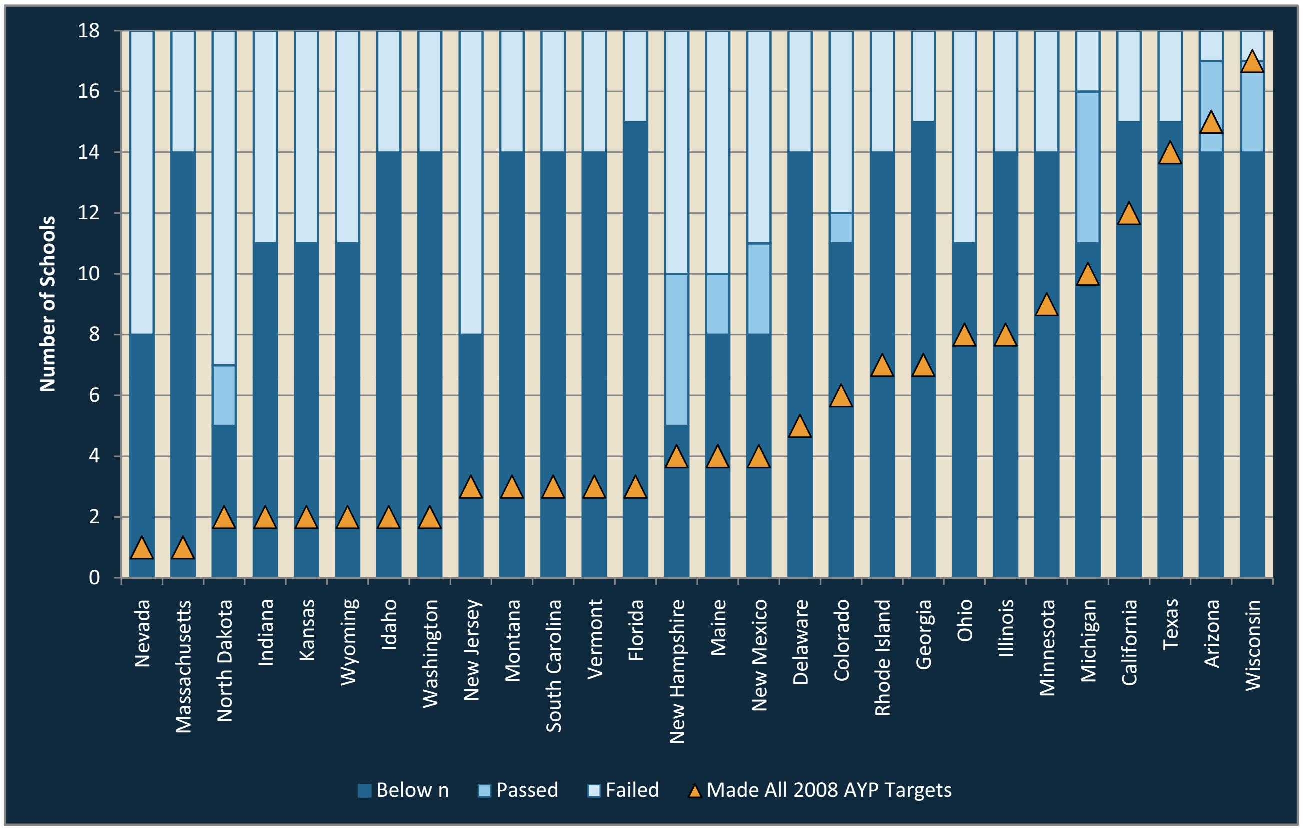

Figure 17. Number of elementary schools in which LEP students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Ohio every elementary school with a qualifying LEP subgroup failed to meet its AMO. In New Hampshire, however, five schools did not meet subgroup requirements and five schools met LEP targets (dark blue and median blue bars). However, even though ten schools met their LEP targets in New Hampshire, only 4 of the 10 schools ultimately made AYP (indicated by the orange triangle). The remaining 6 failed to make AYP because of some other subgroup

consin, all minority subgroups passed in sixteen schools, yet only seven ultimately made AYP.

Performance of LEP students

In general, LEP students are required to participate in state testing for purposes of determining AYP. Students who are not English proficient and are new to the United States need not participate in state testing during the first calendar year in which they're enrolled. Until recently, students who graduated from LEP status by achieving English proficiency were moved out of the subgroup during the year that they became proficient. In practice, this created a churning effect, in which successful students were removed from the LEP subgroup and new English language learners moved in. A mid-course change to NCLB regulations by the U.S. Department of Education now allows states to retain in the LEP subgroup, for up to two years, students who have become proficient in English. This reduces, but does not eliminate, the churning effect.

Many of the elementary schools in the sample (67% of cases) did not have LEP populations large enough to meet

Table 11. Elementary school sample performance relative to their 2008 AMOs for students with limited English proficiency

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which the LEP group was below the minimum subgroup size	336 (67%)
Number of cases in which the LEP group met all AMOs	24 (5%)
Number of cases in which the LEP group failed to meet one or more AMOs	144 (27%)

Note: Percentages may not add to 100 due to rounding.

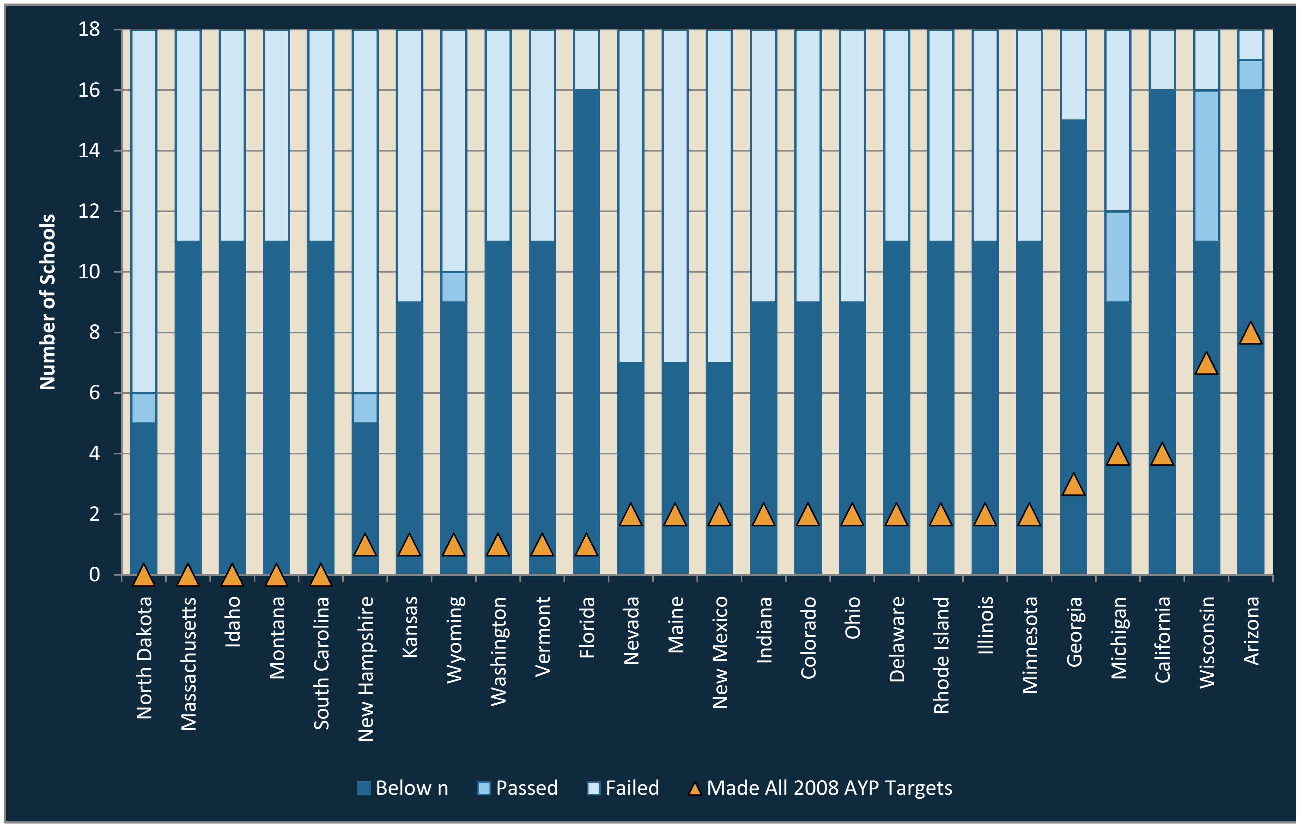

Figure 18. Number of sampled middle schools in which LEP students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. In the vast majority of states (New Mexico, Indiana, Colorado, Delaware, etc.), every school with a qualifying LEP subgroup failed to meet its AMO.

the minimum *n* size in the states studied (Table 11). In situations where this subgroup's performance is counted, however, nearly all schools failed to meet their AMOs. Schools failed in 27% of total cases, nearly six times the number of cases in which schools succeeded (5%). In 20 of the states studied, all schools whose LEP population exceeded the minimum *n* size failed to meet their AMOs (indicated by the absence of a median blue bar in Figure 17).

The middle schools, again, did not perform as well as the elementary schools. Although the majority (57%) did not have LEP subgroups large enough to qualify for evaluation, a school with a qualifying count passed its AMOs in only 3% of the total cases and failed in 40% of the total cases (Table 12). In 20 of the 26 states, all schools with qualifying LEP populations failed to meet their AMOs for this subgroup (Figure 18).

Sadly, the best way to for a school to avoid failure with its LEP students is to avoid having many of them. In fact, more than half of the sample was not evaluated on the performance of these students because they fell below the various states' minimum *n* size requirements (Table 12). And nearly all of those schools that did have a qualifying LEP subgroup failed to meet the AMOs for this group.

Table 12. Middle school sample performance relative to their 2008 AMOs for LEP students

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which the LEP group was below the minimum subgroup size	269 (57%)
Number of cases in which the LEP group met all AMOs	12 (3%)
Number of cases in which the LEP group failed to meet one or more AMOs	187 (40%)

Note: While twenty-eight states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to twenty-six states.

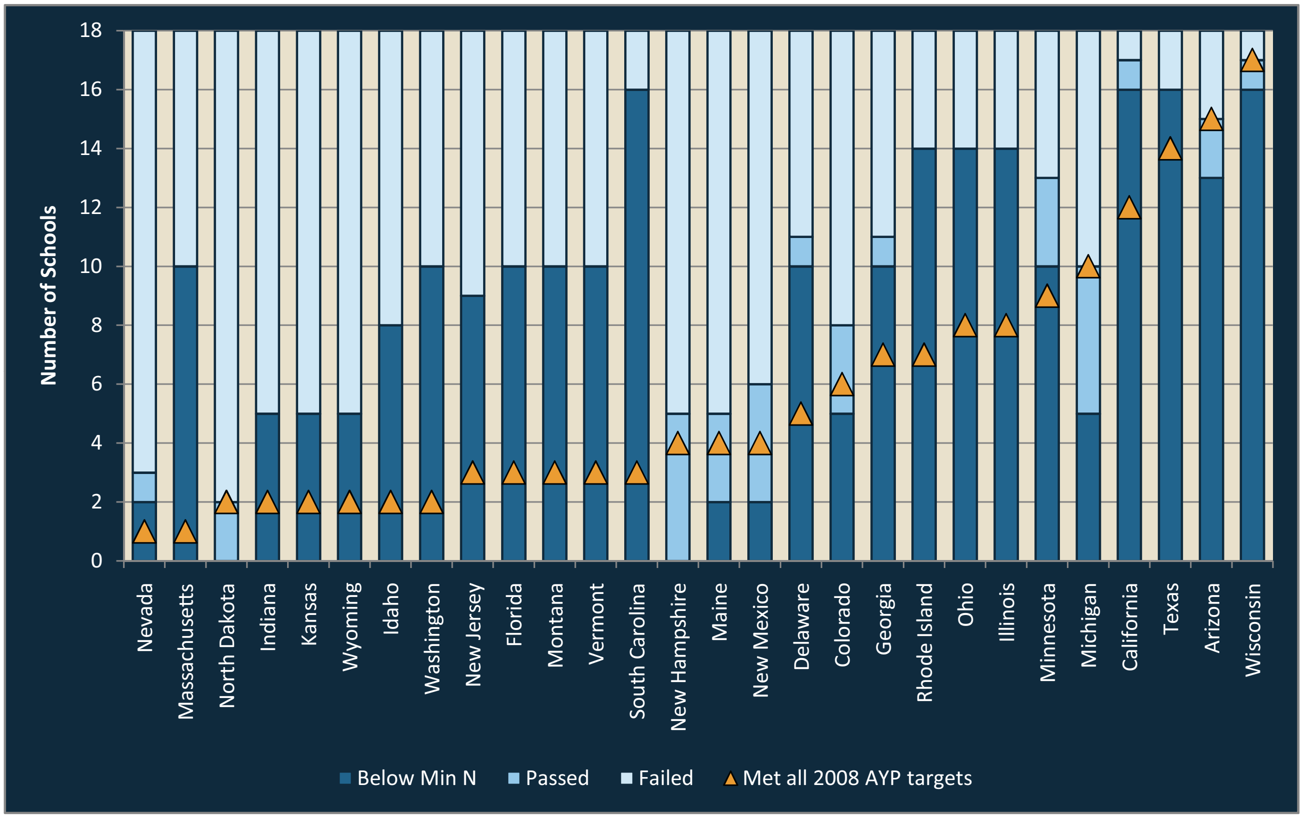

Figure 19. Number of sampled elementary schools in which SWDs met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. In the vast majority of states (Wyoming, Idaho, Washington, Vermont, etc.), every school with a qualifying SWD subgroup failed to meet its AMO.

Performance of SWDs

This was the final factor considered. Students with disabilities are not exempt from the NCLB 100% proficiency requirement, but states are allowed to exclude from testing up to one percent of students who have significant cognitive disabilities. States are also allowed, under a change to the NCLB regulations, to test another two percent of students using an alternative assessment.[4]

Table 13. Elementary school sample performance relative to their 2008 AMOs for students with disabilities

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which the SWD group was below the minimum subgroup size	247 (49%)
Number of cases in which the SWD group met AMOs	32 (6%)
Number of cases in which the SWD group failed to meet one or more AMOs	225 (45%)

How does the SWD subgroup perform? Within the elementary school sample, the count of disabled students fell below the minimum *n* size in just under half of all cases (49%) (Table 13). There were 225 cases of subgroups failing to meet AMOs (45%) and only 32 cases (6%) in which the subgroups met their AMO. In fifteen states, all elementary schools whose SWD subgroup met the required minimum *n* size failed to meet their AMOs (Figure 19).

[4] Participating schools in this study did not report to us whether each student's achievement level was attained on the state's general assessment or on the alternative assessment, so we caution that some students included in these results could be eligible to take a state's alternate assessment or excluded from testing entirely. However, it's not general practice for schools to test students with severe cognitive disabilities on the NWEA assessment, so it is unlikely that these students are included here.

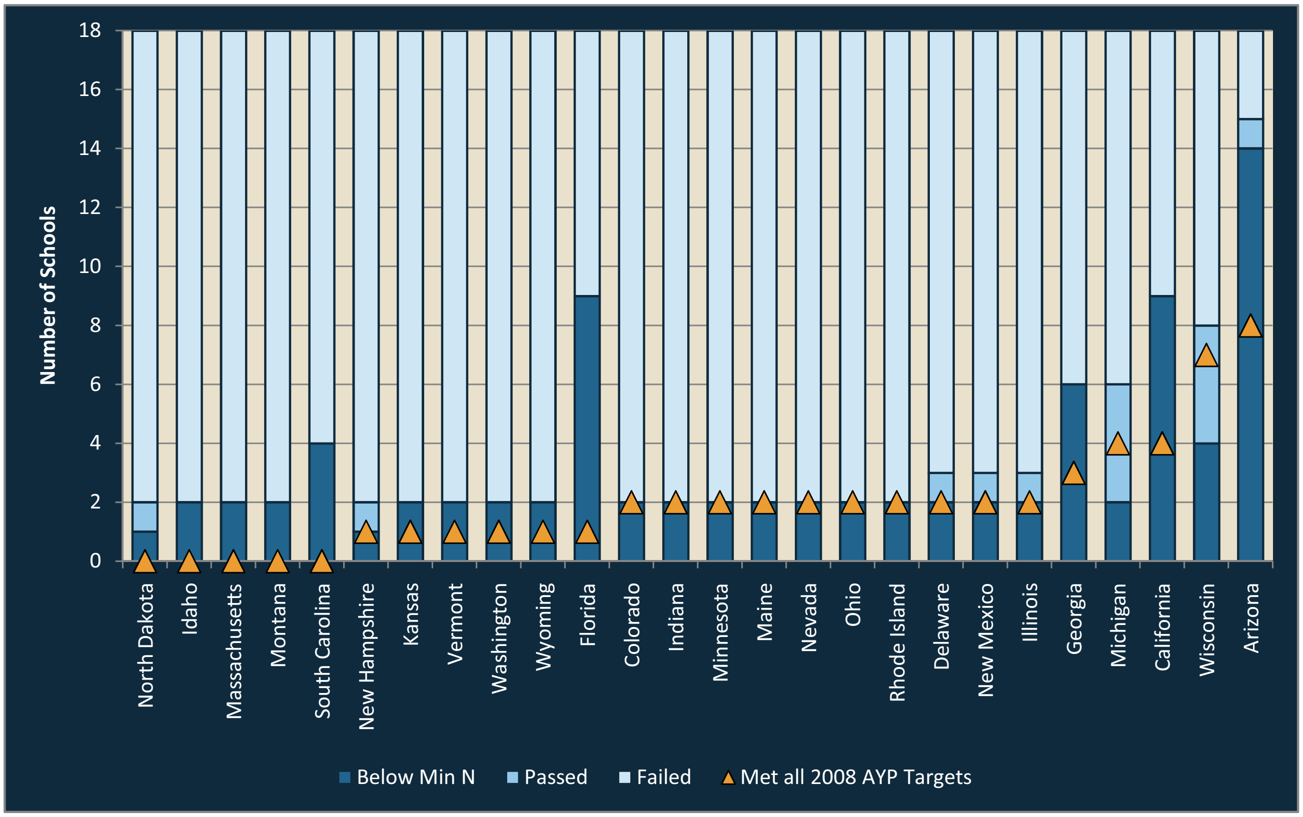

Figure 20. Number of sample middle schools in which SWDs met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. In the vast majority of states (Wyoming, Idaho, Rhode Island, Vermont, etc.), every school with a qualifying SWD subgroup failed to meet its AMO.

Among the middle school sample, in only 18% of cases did schools not have SWD subgroups large enough to qualify for evaluation (Table 14). Of the remaining cases where schools did have large enough SWD subgroups, middle schools met their AMOs in 3% of cases and failed to meet their AMOs in 79% of cases. In 18 of the states, no middle school surpassing the minimum *n* size met its AMO target for SWDs (Figure 20).

Table 14. Performance of the sampled middle schools relative to the 2008 AMOs for SWDs

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which the SWD group was below the minimum subgroup size	84 (18%)
Number of cases in which the SWD group passed AMO	14 (3%)
Number of cases in which the SWD group failed one or more AMOs	370 (79%)

Note: While twenty-eight states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to twenty-six states.

As with LEP students, nearly all of the schools in the sample that have SWD subgroups exceeding the minimum count failed. Because middle schools are generally larger than elementary schools, there are far more cases in which the middle school sample is evaluated (82%) than in the elementary schools (51%).

The Lowdown on Subgroup Performance

Figure 21 provides a very interesting summary of how subgroup performance affects the prospects for making AYP within our sample. Essentially it shows that schools had much more success with their low-income and minority subgroups than with their LEP and SWD subgroups. The graphic also shows that elementary schools

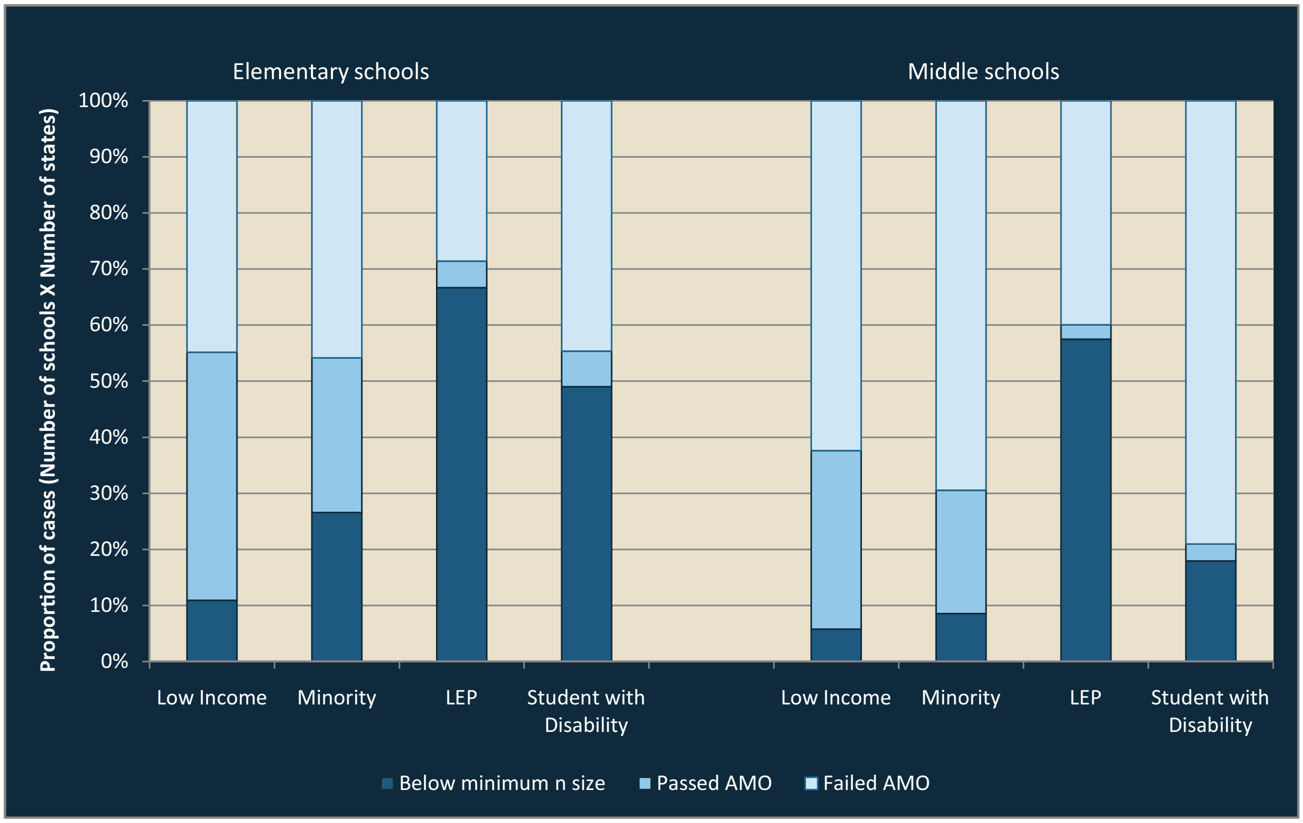

Figure 21. Summary of subgroup performance relative to AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. The figure shows that schools had much more success with their low-income and minority subgroups than with their LEP and SWD subgroups. It also shows that elementary schools failed to meet their AMOs with far less frequency than middle schools, primarily because elementary schools had far fewer subgroups that met the minimum subgroup size.

Abbreviations: SWDs = students with disabilities; AMO = annual measurable objective (yearly target)

failed their AMOs with far less frequency than middle schools, primarily because elementary schools had far fewer subgroups that met the minimum subgroup size.

While the low passing rates of low-income and minority subgroups may be frustrating, the passing rates for schools with qualifying LEP or SWD subgroups are simply astounding (as shown by the sliver of median blue in these categories in Figure 21). In the vast majority of cases, a school with a qualifying subgroup in one of these two categories failed to meet the relevant AMOs and thus failed to make AYP.[5] **The difficulty of the states' cut scores and AMOs were largely irrelevant in these cases. These subgroups failed whether the cut scores were high or low and whether the AMOs were strict or generous.**

So, to summarize:

- A state's minimum subgroup size (or *n* size) determines the number of subgroups that must meet an AMO. Since failing a single AMO causes a school to fail to make AYP, having more subgroups increases the number of opportunities for failure. This is the case with middle schools in the sample—they don't fare worse because they are less effective in educating students, but because they have more subgroups.

[5] We should note that this study may underestimate the performance of students in the LEP and SWD subgroups, mostly because of the likely differences between how LEP students and SWDs are treated in MAP, the assessment we used in this study, and in the various state standardized tests. Specifically, the U.S. Department of Education has issued new NCLB guidelines in recent years that exclude small percentages of LEP students and SWDs from taking the state test or that allow them to take alternative assessments. In this study, however, no valid MAP scores were omitted from consideration.

- Rather than claim that large schools face a "diversity penalty," it may be fairer to say that small schools enjoy a "homogeneity bonus." Small schools typically do not have to meet objectives for many subgroups since they don't have enough low income, minority, LEP or SWD students to qualify for evaluation. In large schools, these subgroups often fail to meet their AMOs (as shown in Figure 21). Because there's no reason to believe that pupils in small-school subgroups are performing at levels way beyond those in larger-school subgroups, small schools are probably fortunate that they're not accountable for these groups separately. They clearly have an easier time making AYP than larger schools.

- As indicated above, middle schools in the sample fared more poorly than elementary schools. In only 32% of cases did low-income student subgroups in middle schools meet their AMOs. Contrast this with elementary schools, where 44% of low-income subgroups met their AMOs. The picture is much the same for minority subgroups. In 22% of middle school cases, all minority student subgroups met their AMOs; the same is true in 28% of elementary school cases.

- Even more damaging to a school's chances of making AYP is the presence of a qualifying subgroup of LEP students or SWDs. In only 3% of middle school cases and 5% of elementary school cases did a LEP subgroup meet its AMOs. Similarly, in only 3% of middle school cases and 6% of elementary school cases did a subgroup of SWDS meet its AMOs. As a result, most schools that actually made AYP by our estimate did so because their LEP and SWD subgroups were too small to qualify for evaluation.

Limitations

The purpose of this study was to explore how key elements of NCLB, in this case proficiency cut scores, proficiency rate targets (AMOs), subgroup sizes, and confidence intervals may interact to affect the AYP status of schools. We hoped to shed light on such questions as "Would a school with a population and performance mix that makes AYP in California also be likely to make AYP in New Hampshire, Washington, or South Carolina?"

A sample of real schools was chosen for the study in an effort to assure a meaningful connection between our analysis and the actual conditions faced by schools. (Each school is identified by a pseudonym.) We hope this makes the study useful, informative, and interesting. This study literally shows what happens when you take the performance of a set of schools on a single assessment, estimate different proficiency cut scores for that assessment based on a sound estimate of the difficulty of the standards in different states, and apply the AYP rules in place for that state to the dataset. This kind of illustration is very useful when one wants to evaluate whether the effect of the NCLB accountability policy is likely to be consistent across states. And that was our purpose here.

We must emphasize, however, that the MAP assessment and analytic tools will not precisely replicate the sample schools' performance on their state tests. While all students in the sample took some form of their state assessment, schools did not identify whether students took the regular assessment or the alternative assessment. For the purposes of our study, a student's performance on the various states' assessments was projected from their MAP scores. Therefore, it is possible that some students we identify as failing, particularly LEP students or students with disabilities, would be eligible to take the alternative form of the assessment in some states. We have no data that allow us to predict how these students might have performed on the alternative assessment.

Some students within a school who participated in state testing did not participate in MAP testing (and vice versa), but we included only students who participated in both MAP and state tests in our sample. As a result, the students included for estimation in our study were not identical to the students who participated in state testing that same school year. Tables A-4 and A-5 (in Appendix A) show differences in the count of students taking MAP and their state test and those who participated only in their state test for the sample schools. For all but two of the sample schools, the MAP results predicted, within five percentage points, the school's actual performance on their state test. In addition, our pilot study (Cronin et al. 2007b) found that the rates of proficiency estimated on the MAP assessment for samples of students closely paralleled the rates of proficiency reported on state tests.

In testing the effects of confidence intervals, we followed the methodology employed by the state in their calculations. Because MAP is an adaptive assessment[6] (state tests are generally fixed form), our estimate of the confidence intervals associated with MAP may be narrower in some states than the confidence interval associated with the state assessment. This happens because the standard error of measure associated with MAP is generally smaller for very high and low performing students than the standard error of measure on a fixed form test. In these circumstances, our confidence interval calculation may slightly understate the actual effect of the confidence interval within that state.

In addition, certain conditions used by states to determine AYP status were not evaluated as part of this study. Some schools identified in our illustration as failing to make AYP would make it because they met their state's safe harbor provisions. Some would now also pass under the growth-model pilot underway in a handful of states, such as Ohio. In this respect, our findings do underestimate the actual AYP performance of some of the schools in the sample. Conversely, a few schools identified as making AYP might actually fail to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of a particular subgroup(s) within their student population. While we concede that our results may understate actual AYP performance in some cases, we believe the study provides a relatively accurate and useful prediction of how schools generally fare under the *base* AYP rules. That is, if NCLB was intended to get 100% of students, including those within subgroups, across the proficiency bar, the study illustrates how well the sample schools fared relative to this goal and its benchmarks.

With these limitations considered, we believe this study illuminates the inconsistency of AMOs and proficiency cut scores and other rules for determining AYP status across states. It does not, however, necessarily replicate with precision the performance and AYP status of the sample schools within their own state, or predict with complete consistency their status if students took the exams required by other states.

[6] This means that students are offered questions at a level of difficulty that reflect their current performance rather than their current grade. For example, a high-performing third-grader might receive questions at the fifth-grade level, while her lower-performing peer might receive questions pegged at the first-grade level.

DISCUSSION

NCLB was intended to ensure that all schools set high standards for reading and math, and to hold all students accountable to these standards, regardless of their ethnicity, income, or other differences. Unfortunately, the strategy chosen to implement these goals creates an illusion of accountability that will not get us to these results, in part because it was too lax in establishing guidelines around standards and rules and too inflexible in its requirements for outcomes.

NCLB has given states the discretion to establish proficiency cut scores, the required trajectory for improvement, minimum subgroup sizes, and confidence intervals. Our results show that the product of these differences bears no resemblance to a coherent system. Not only do the proficiency cut scores themselves vary greatly, but the variance in improvement trajectories, subgroup sizes, and policies for application of confidence intervals result in wildly different Adequate Yearly Progress results for the schools in our sample. It appears, then, that the federal government has implemented a system in which geography had as much to do with our schools' AYP status as their students' academic performance. In addition, it was sometimes impossible to distinguish between the high-performing and underperforming schools in our sample. We could argue that NCLB has been too lax in allowing this degree of discretion.

Conversely, the law requires 100% of students, including 100% of students in every subgroup, to achieve the states' proficiency standards by 2014. In the meantime, each and every subgroup is required to meet the Annual Measured Objectives that are set for schools each year. These subgroups include low-income students and ethnic minorities, but they also include subgroups whose members have documented academic challenges, such as Limited English Proficient students and Students with Disabilities students and SWDs. Although the sample schools in the study met proficiency goals for their overall student populations in the majority of cases, the performance of subgroups within the sample schools was far worse. All eligible minority subgroups within a school met their proficiency objectives in only 20% to 30% of cases. But eligible LEP and SWD populations fared even worse. Within the sample schools, these two groups met their proficiency objectives in just 3% to 6% of cases. This means that the relative difficulty of the cut scores and the AMOs are essentially irrelevant, because LEP and SWD subgroups failed even in states with low cut scores and AMOs. In this regard, we could argue that NCLB has been too strict.[1]

Of course the bottom line for schools is whether they ultimately make AYP. Applying these rules to the elementary sample, we found that AYP results differed dramatically across the states studied. The number of schools in the sample that made AYP varied from 1 in Massachusetts and Nevada to 17 in Wisconsin. Ultimately there was no consistency in the way elementary schools were judged, meaning that there is likely to be no consistency in the way sanctions are applied.

The results for the middle school sample were consistent but grim. In 5 states none of the schools in the sample met AYP; in 6 other states, only 1 school made AYP. In general, the higher rates of failure can be attributed to the fact that middle schools were accountable for more subgroups. In many cases, the failing subgroups were low-income students and ethnic minorities. But in almost all cases in which the school was accountable for a LEP or SWD subgroup, the school failed.

We could take this to mean that the AYP fate of many schools is tied to the performance of their lowest per-

[1] It's important to note that federal reports regarding SWD and LEP subgroup performance differ from our findings here. The National Assessment of Title I: Interim Report (2006) concluded that 23% of schools (they were not broken down by elementary and middle) failed to make AYP in 2003-2004 due to the performance of a single subgroup. Of this 23%, the breakdown was as follows: 13% of schools missed AYP due to the performance of students with disabilities, 4% because of LEP performance, 3% because of low- income student performance, and 3% because of the performance of a single ethnic group. The differences between the federal report and this one may be due to several factors, including: (1) the relatively new NCLB guidelines that exclude small percentages of LEP students and SWDs from taking the state test or that allow them to take alternative assessments; (2) the fact that this report does not calculate the impact of safe harbor on subgroup performance; and (3) the study sample is not nationally representative.

forming subgroup, frequently a subgroup with documented learning challenges. From our results, we could also extrapolate that a school's best strategy for making AYP would be to rid itself of the LEP and SWD subgroups because the presence of one essentially guarantees failure, even in circumstances where these two subgroups outperform similarly identified students in other schools. If that's truly the case, it's unlikely that the current handling of subgroups within NCLB is likely to improve the results achieved.

Some might conclude that we're arguing for different or lower proficiency standards—or both—for LEP students and SWDs. Let's be clear: That's not our argument at all. Instead, we believe the evidence shows that evaluating schools primarily on whether their students meet a fixed, arbitrary, and often low proficiency bar serves all students poorly, including LEP students and SWDs. After all, these students are not members of a homogenous subgroup. LEP students may include some who enter the United States in their teenage years with no formal schooling alongside others who may have attended elite private schools abroad and have exposure to multiple languages. SWDs can range from learners who are academically gifted but challenged by dyslexia, those who perform below their ability because they have behavioral issues, and those with significant cognitive barriers that make learning slower and more difficult. How well is a gifted, dyslexic learner served by meeting a standard that's set to the least-common denominator of performance? And what about a student in Massachusetts (a state with high standards and difficult targets) who has shown promising growth despite huge learning difficulties, but has not yet achieved proficiency? Is that student served well if her school is sanctioned because she and some of her peers did not all achieve a standard that's set to college readiness?

We strongly believe that parents should know how their child is progressing relative to their family's aspirations (which are almost always college readiness). But checking off the number of students who cross a fixed—and low—proficiency bar is a poor way to judge school effectiveness. We believe students would be better served by a model that focuses on how effective schools are in promoting student growth. Such a model would require schools to focus their energy on all students—high-, average-, and low-performing—as well as members of subgroups, which could only be beneficial to both school and student. And a model like this would keep schools from focusing all of their energy on the relatively few students who have the best prospects for crossing a proficiency bar during the current year.

On a technical note, the use of confidence intervals seems to have emerged as a coping mechanism for some of NCLB's design problems. Ostensibly the confidence interval exists to account for the possibility of some form of measurement error in the performance of the student population. In 8% to 11% of cases, a school that wouldn't have met the AMOs for overall proficiency ended up meeting its target with the assistance of a confidence interval. We included (but did not report) the confidence interval in the calculation of subgroup performance as well. There is no doubt that the confidence interval helps many subgroups meet their AMOs, subgroups that wouldn't have otherwise met these targets. But the fact that the vast majority of schools (particularly among our middle school sample) still ultimately failed to make AYP suggests that the confidence interval was not the "difference maker" with many schools. That said, we think the logic for including confidence intervals in NCLB's accountability system is weak, and we doubt confidence intervals would be required in a more consistent, rational accountability system.

Taken as a whole, the evidence from the sample suggests that NCLB, as currently implemented, is not a discriminating system. A tremendous amount of money and energy has been spent to create the illusion of accountability. But the accountability is not coherent. We found states where most schools failed to make AYP and others where nearly every school made it. We found demonstrably good schools that failed AYP far too often, and some pretty mediocre schools that slid by in some states. So in reality, what passes for accountability feels more like a high-dollar crapshoot. Some schools may really be failing—no doubt that's so—but they get off easy. For others, the dice aren't as kind—they get labeled as failing but are truly competent.

Either way this is not the type of accountability that will, in the long run, really improve schools, states, or nations.

APPENDIX A:

COMPLETE METHODOLOGY

The purpose of this study was to explore how key elements of NCLB, in this case proficiency cut scores, proficiency rate targets, subgroup sizes, and confidence intervals, interact to affect the AYP status of schools. We hoped to shed light on such questions as "Would a school with a population and performance mix that makes AYP in California also be likely to make AYP in New Hampshire, Washington, or South Carolina?" We pursued this by applying each state's proficiency cut scores and several key rules related to AYP to achievement data from a multistate sample of schools that were chosen to reflect a broad range of student performance, income, and growth in student achievement.

Sample

We started by creating two samples. The first was a sample of states for which we could compare cut scores and AYP rules. The second was a sample of schools for which we could use achievement data to evaluate the impact of the various cut scores and rules on their possible AYP status.

States Sample

In all, we included 28 states in our study (see Table A-1). States were included in the study if sufficient student records from state and NWEA testing were available to permit a robust estimate of the state's proficiency cut scores in both reading and math for grades three through eight.[2] Twenty-six of these cut score estimates were originally reported in *The Proficiency Illusion* (Cronin et al. 2007a). To estimate the majority of cut scores used in this study, we used achievement data from the 2005–2006 school year. Since *The Proficiency Illusion* was published, cut scores for 3 additional states were estimated using achievement data from the 2006–2007 school year. Cut scores were estimated for grades three through eight, and these were used to determine the proficiency rates of the sample schools. There were some exceptions, as follows:

Table A-1. States and grades included in the study sample and terms used for alignment estimate*

State	Term	Grades †
Arizona	Spring 2005	3,4,5,6,7,8
California	Spring 2006	3,4,5,6,7,8
Colorado	Spring 2005	3,4,5,6,7,8
Delaware	Spring 2006	3,4,5,6,7,8
Florida	Spring 2007	3,4,5,6,7,8
Georgia	Spring 2007	3,4,5,6,7,8
Idaho	Spring 2006	3,4,5,6,7,8
Illinois	Spring 2006	3,4,5,6,7,8
Indiana	Fall 2006	3,4,5,6,7,8
Kansas	Fall 2006	3,4,5,6,7,8
Maine	Spring 2006	3,4,5,6,7,8
Massachusetts	Spring 2006	3,4,5,6,7,8
Michigan	Fall 2005	3,4,5,6,7,8
Minnesota	Spring 2006	3,4,5,6,7,8
Montana	Spring 2006	3,4,5,6,7,8
Nevada	Spring 2006	3,4,5,6,7,8
New Hampshire	Fall 2005	3,4,5,6,7,8
New Jersey††	Spring 2006	3,4,5,6,7
New Mexico	Spring 2006	3,4,5,6,7,8
North Dakota	Fall 2006	3,4,5,6,7,8
Ohio	Spring 2007	3,4,5,6,7,8
Rhode Island**	Fall 2005	
South Carolina	Spring 2006	3,4,5,6,7,8
Texas††	Spring 2006	3,4,5,6,7
Vermont	Fall 2005	
Washington	Spring 2006	3,4,5,6,7,8
Wisconsin	Fall 2005	3,4,5,6,7,8
Wyoming	Spring 2007	3,4,5,6,7,8

*The table shows that a number of states administer their state assessment in the fall. For these states we estimate the cut score using fall data and convert that estimate to the equivalent spring score, using percentile ranks. This permits us to evaluate each state's results using NWEA data from a single term.

** Rhode Island, Vermont, and New Hampshire use the New England Common Assessment Program Tests. Cut score estimates for these states are based on the estimates for New Hampshire.

†The same grades were included for both math and reading.

††Because eighth-grade cut scores for New Jersey and Texas couldn't be estimated, we didn't include these states in the middle school portion of the study.

[2] We require a sample of 700 or more students at each grade to generate a cut score estimate.

- New Hampshire, Rhode Island, and Vermont report on AYP using a common, jointly developed state test called the New England Common Assessment Program (or NECAP), and all three states use the same proficiency cut scores on that test to evaluate student performance. The rules used to evaluate school AYP, though, including annual targets, differ across the three states. Our estimated cut scores on NECAP were derived from a sample of New Hampshire students, but our AYP analyses consider each state's rules separately.

- No school districts within Maryland use NWEA tests for math, so cut score estimates were available only for reading. Consequently, although Maryland reading cut scores were reported in *The Proficiency Illusion*, Maryland is not included in the current study.

- Sample sizes were inadequate to produce eighth grade cut score estimates in Texas and New Jersey. In these cases, we analyzed only elementary schools under the AYP rules in these two states.

Example Schools

We chose 36 schools to serve as example schools in the study, treating the data from students in these schools as if the school existed in each of the 28 sample states (26 for middle schools). We designed the school selection process to produce a group of schools that reflected breadth in student achievement, school size, diversity, and student growth. The selected schools do not necessarily reflect the demographics of the nation as a whole, nor was that our intention. To create the sample, we contacted 20 school systems to request their participation in the study. Eight school systems that included 153 district and charter schools in the states of Arizona, California, Illinois, Kansas, South Carolina, Washington, and Wisconsin agreed to participate. These school systems supplied student demographic data and state test results to supplement NWEA achievement data that were already stored. Of these schools, 103 were elementary schools and 50 were middle schools. Before we selected the schools, we compiled data on each, relative to the following variables:

- **Student performance (net student achievement in reading and math):** The average raw scale score difference between the students' performance and the median performance (based on NWEA [2005]) for their grade in this subject. As a rule of thumb, a difference of six scale score points is roughly equivalent to a difference of one school year in median achievement.

- **Income level (proportion of school population eligible for free or reduced-price lunch):** This was the only available variable that is a surrogate for family income.

- **Student grade (elementary and middle school groupings):** One finding from *The Proficiency Illusion* (Cronin et al. 2007a) was that middle schools tended to have more difficult standards than elementary schools relative to the NWEA norms. In addition, some states set different AMOs (percentages of students required to meet standards) for elementary and middle school grades. Finally, middle schools, on average, enroll more students than elementary schools. As a result, we created two study groups one composed entirely of middle schools, the other comprising only elementary schools.

- **Student growth (net student growth in reading and math):** This is the average scale score difference on NWEA's assessment, the Measures of Academic Progress (MAP) between student scores in fall to spring terms relative to the NWEA RIT Point Norms (NWEA 2005). This metric compares the average growth of students to the growth of students who started with the same scale score in that grade.

Within the elementary and middle school groups, we ranked and classified schools relative to their peers on the achievement, income, and growth variables. Three categories (high, middle, low) were created for the student achievement and student growth variables, with the upper third of schools assigned a classification of high, the middle third assigned average, and the bottom third assigned low. For the income classification, we created two categories. The fifty percent of schools with the highest free or reduced-price lunch population were classified as low income, the other half as high income.

Next, we compiled these classifications into a code that described the achievement, income, and growth status of each school. Thus, a school classified as *high, high, low* (HHL) would be classified as high-achieving, high-income, and low-growth. Eighteen codes were possible (3 achievement × 3 growth × 2 income).

In addition to selecting schools that reflected diversity on these three criteria, we also attempted to select schools in which student performance on their respective state tests was closely predicted by the NWEA assessment. Accordingly, we tried to find schools in which the estimated proficiency rate of students in both reading and math on the NWEA test was within 5% of their actual proficiency rate on their particular state's test.

Here are details of the process we used to select schools:

1. For each cell (e.g., high-achievement, high-income, high-growth), we attempted to find one or more schools with that cell assignment. If there was no school with that cell assignment, we attempted to assign a school with an adjacent assignment, proceeding in the following order (growth → achievement → income). Tables A-2 and A-3 present the results of the sample schools relative to these criteria.

2. Once one or more schools were identified, we selected schools whose predicted proficiency rate on both the NWEA reading and math assessment was within 5% of the actual proficiency rate attained by the school on their own state test. If more than one school met this criterion, we randomly selected a school. If no school met this criterion, we attempted to find a school that met the criterion from an adjacent cell. Tables A-4 and A-5 report the performance of the sample schools on these criteria.

3. In circumstances in which no school met the requirement for predicted proficiency, we selected the school whose actual state test performance was most closely predicted by the NWEA assessment.

The names of the schools selected were changed to protect their anonymity. We also altered the state and school type for Barringer School, whose identity might be discerned from the school's size and unusual configuration if its state and school type were known.

The data indicate that the elementary schools as a group showed slightly higher than average student performance and slightly higher than average growth when compared with students in NWEA's norming group as a whole (NWEA 2005). The average performance of the middle school group was also higher than the norming group, although the growth of these students was slightly below average. Because the study group had slightly higher than average performance, this group might achieve higher rates of proficiency than a group of schools randomly selected from NWEA's 2005 norming population.

In constructing our sample, we didn't aggregate any information that would communicate the projected proficiency rate of students (on the NWEA test) or the actual size of any subgroup within a school, with the exception of the free and reduced-price lunch rate. We did this intentionally to ensure that the selection process was as free as possible from bias that might derive from having direct knowledge of how the school might fare under the AYP rules of any given state. For example, if we had known that one of the selected schools had 41 Hispanic/Latino students, we would also know that this particular subgroup would be large enough to require AYP consideration in some states but not others. Not compiling this kind of information in advance helped to ensure that the schools—although selected purposefully—were not chosen with knowledge that a school's selection would produce a predetermined result in the various states.

Table A-2. Status of elementary school study group on the selection variables*

Pseudonym	State	Type	State tested in Math	NWEA performance†	Income (percentage in parentheses)	Performance (percentage of average growth in parentheses)‡	Assigned category§	Actual category
King Richard	Illinois	District elementary	415	High (+10.1)	High (13)	High (140)	HHH	HHH
Roosevelt	Wisconsin	District magnet (gifted)	284	High (+8.9)	High (13)	Middle (103)	HHL	HHM**
Marigold	Illinois	District elementary	372	High (+7.7)	High (17)	Middle (122)	HHM	HHM
Forest Lake	South Carolina	District elementary	378	High (+7.6)	High (34)	High (152)	HLH	HHH**
Paramount	Arizona	District elementary	270	Middle (+4.2)	High (37)	Middle (142)	MHM	MHM
Coastal Intermediate	South Carolina	District intermediate	550	Middle (+3.8)	Low (58)	High (131)	MLH	MLH
Winchester	California	District elementary	262	Middle (+3.5)	High (13)	High (139)	LHH	MHH**
Wayne Fine Arts	Wisconsin	District alternative	168	Middle (+2.4)	High (22)	Low (97)	MHL	MHL
Alice Mayberry	South Carolina	District elementary	295	Middle (+2.0)	Low (60)	Low (88)	HLL	MLL**
Wolf Creek	California	District elementary	281	Middle (+0.6)	High (25)	High (133)	MHH	MHH
Scholls	South Carolina	District elementary	279	Middle (+.06)	Low (61)	High (111)	MLM	MLH
Hissmore	South Carolina	District elementary	274	Middle (+0.6)	Low (75)	Middle (103)	MLM	MLM
Island Grove	Washington	District elementary	280	Middle (–2.5)	High (40)	Middle (117)	LHM	MHM**
John F. Kennedy	South Carolina	District elementary	268	Middle (–2.0)	Low (75)	Low (94)	MLL	MLL
Nemo	Wisconsin	District elementary	188	Middle (–2.8)	High (33)	Low (93)	LHL	MHL**
Few	Arizona	District elementary	263	Low (–6.0)	Low (90)	High (135)	LLH	LLH
Maryweather	Arizona	District elementary	224	Low (–7.1)	Low (80)	Middle (113)	LLM	LLM
Clarkson	California	District elementary	434	Low (9.4)	Low (87)	Low (55)	LLL	LLL

*Group is sorted by math and reading performance. Within the table, H stands for high, M for middle, and L for low.

†The number in parentheses reflects the average scale score difference in performance and growth (in math and reading) between students in the school and those in the norming group.

‡The number in parentheses represents the average scale score improvement shown by this school relative to a matched group of students from the NWEA norming group. One hundred percent means that a school is on target in terms of expected growth. Less than 100% growth means that the average student is increasing *by less* than normative amounts, while percentages over 100 mean that the average student is *exceeding* normative growth expectations.

§Performance/income/growth

**Indicates that the selection was from an adjacent cell

Table A-3. Status of middle school study group on the selection variables*

Pseudonym	State	Type	State tested in math	NWEA performance†	Income (percentage in parentheses)	Performance (percentage of average growth in parentheses) ‡	Assigned category§	Actual category
Chaucer	California	District middle	1083	High (+10.4)	High (10%)	High (175%)	HHH	HHH
Walter Jones	Arizona	District magnet	165	High (+6.5)	High (38%)	Middle (111%)	HLH	HHM**
Artemus	Illinois	District middle	749	High (+5.8)	High (17%)	Middle (91%)	HHM	HHM
Ocean View	California	District middle	599	High (+3.6)	High (22%)	Middle (85%)	HHL	HHM**
Zeus	South Carolina	District middle	947	Middle (+2.2)	High (42%)	Middle (85%)	MHL	MHM**
Lake Joseph	Washington	District middle	801	Middle (+1.8)	High (34%)	High (111%)	LHH	MHH**
Black Lake	South Carolina	District middle	1380	Middle (+1.7)	Low (46%)	Middle (87%)	HLM	MLM**
Hoyt	South Carolina	District middle	1012	Middle (+0.8)	Low (55%)	Low (79%)	HLL	MLL**
Kekata	South Carolina	District middle	885	Middle (+0.5)	Low (57%)	Middle (103%)	MLM	MLM
Barbanti	California	District middle	1459	Middle (-0.6)	High (45%)	High (130%)	MHH	MHH
Filmore	Washington	District middle	674	Middle (-0.7)	High (40%)	Middle (96%)	MHM	MHM
Chesterfield	South Carolina	District middle	539	Middle (-2.4)	Low (63%)	Low (75%)	MLL	MLL
Tigerbear	South Carolina	District middle	702	Middle (-3.4)	Low (78%)	Middle (87%)	MLH	MLM**
McCord	Wisconsin	Charter	730	Low (-3.7)	High (41%)	Middle (95%)	LHL	LHM**
Pogesto	Washington	District intervention	83	Low (-3.9)	Low (46%)	Middle (107%)	LLH	LLM**
Barringer (K-8)	***	***	2198	Low (-5.0)	Low (81%)	Low (77%)	LLL	LLL
ML Andrew	Wisconsin	District middle	651	Low (-5.3)	High (37%)	Middle (85%)	LHM	LHM
McBeal	Arizona	District middle	808	Low (-6.7)	Low (58%)	Middle (87%)	LLM	LLM

*Group is sorted by math and reading performance. Within the table, H stands for high, M for middle, and L for low.

†The number in parentheses reflects the average scale score difference in performance and growth (in math and reading) between students in the school and those in the norming group.

‡The number in parentheses represents the average scale score improvement shown by this school relative to a matched group of students from the NWEA norming group. One hundred percent means that a school is on target in terms of expected growth. Less than 100% growth means that the average student is increasing by less than normative amounts, while percentages over 100 mean that the average student is exceeding normative growth expectations.

§Performance/income/growth

**Indicates that the selection was from an adjacent cell

*** Because of the school's very large student population, the state and type was removed to preserve its anonymity.

Table A-4. Comparison of sampled elementary schools' actual state test performance to estimated performance on NWEA test

Pseudonym	State	State math	NWEA math	Count	State proficiency rate, %		NWEA proficiency rate. %		Difference, %	
		Count	Count	Difference, %	Math	Reading	Math	Reading	Math	Reading
King Richard	Illinois	415	296	29	95.3	89.8	95.6	89.1	–0.3	0.7
Roosevelt	Wisconsin	284	297	–5	93.3	96.3	94.9	98.3	–1.6	–2.0
Marigold	Illinois	372	278	25	94.4	91.0	96.0	86.3	–1.6	4.7
Forest Lake	South Carolina	378	373	1	69.3	69.5	68.1	69.1	1.2	0.4
Nemo	Wisconsin	188	215	–14	65.9	81.3	69.3	85.1	–3.4	–3.8
Few	Arizona	263	291	–11	75.0	54.3	70.8	59.1	4.2	–4.8
Maryweather	Arizona	224	219	2	58.6	51.1	63.5	54.8	–4.9	–3.7
Clarkson	California	435	356	18	19.8	32.4	18.6	32.3	1.2	0.1
Wolf Creek	California	281	218	22	60.9	54.8	57.8	54.8	3.1	0.0
Winchester	California	262	212	19	59.9	58.2	64.2	63.0	–4.3	–4.8
Wayne Fine Arts	Wisconsin	168	174	–4	79.2	92.3	84.0	95.0	–4.8	–2.7
Paramount	Arizona	270	269	0	84.3	79.7	83.0	80.7	1.3	–1.0
Scholls	South Carolina	279	268	4	44.9	48.5	48.0	44.8	–3.1	3.7
Coastal Intermediate	South Carolina	550	523	5	60.7	49.9	57.2	52.1	3.5	–2.2
Island Grove	Washington	280	238	15	58.9	71.6	58.8	71.2	0.1	0.4
Alice Mayberry	South Carolina	295	290	2	46.4	48.6	43.4	47.8	3.0	0.8
John F. Kennedy	South Carolina	268	269	0	33.3	40.8	32.7	38.1	0.6	2.7
Clarkson	California	274	263	4	37.6	46.6	41.4	47.3	–3.8	–0.7

🟧 Note: Light peach shading indicates a greater than 10% difference in the percentage of students tested.

Table A-5. Comparison of sampled middle schools' actual state test performance to estimated performance on NWEA test

Pseudonym	State	State math	NWEA math	Count	State proficiency rate, %		NWEA proficiency rate, %		Difference, %	
		Count	Count	Difference, %	Math	Reading	Math	Reading	Math	Reading
Chaucer	California	1083	1118	-3%	67.8%	68.8%	69.5%	73.5%	-1.7%	-4.7%
Ocean View	California	599	626	-5%	58.7%	63.8%	52.1%	63.6%	6.6%	0.2%
Artemus	Illinois	749	426	43%	89.5%	86.7%	92.0%	82.4%	-2.5%	4.3%
Walter Jones	Arizona	165	172	-4%	87.0%	89.3%	85.5%	85.7%	1.5%	3.6%
Zeus	South Carolina	1018	947	7%	42.6%	41.3%	46.7%	39.9%	4.1%	1.4%
ML Andrew	Wisconsin	651	746	-15%	67.6%	75.6%	71.0%	82.2%	-3.4%	-6.6%
Barringer Charter (K-8)	Illinois	2198	2463	-12%	73.5%	64.1%	76.2%	63.2%	-2.7%	0.9%
Pogesto	Washington	83	54	35%	27.7%	52.3%	31.5%	53.7%	-3.8%	-1.4%
McCain	Arizona	808	888	-10%	53.0%	58.7%	56.0%	59.2%	-3.0%	-0.5%
Barbanti	California	1459	1430	2%	43.8%	45.5%	42.9%	45.3%	0.9%	0.2%
Filmore	Washington	674	584	13%	42.2%	63.9%	46.2%	60.2%	-4.0%	3.7%
McCord Charter	Wisconsin	730	790	-8%	65.8%	78.0%	71.4%	83.2%	-5.6%	-5.2%
Chesterfield	South Carolina	539	523	3%	35.1%	28.7%	30.2%	25.8%	4.9%	2.9%
Hoyt	South Carolina	1012	975	4%	35.1%	31.4%	36.9%	36.0%	-1.8%	-4.6%
Kekata	South Carolina	885	855	3%	39.6%	35.7%	42.6%	35.3%	-3.0%	0.4%
Black Lake	South Carolina	1380	1310	5%	45.0%	35.0%	45.6%	32.8%	-0.6%	2.2%
Tigerbear	South Carolina	702	676	4%	30.6%	25.9%	32.1%	27.3%	-1.5%	-1.4%
Lake Joseph	Washington	801	695	13%	48.4%	68.1%	54.8%	67.3%	-6.4%	0.8%

▨ "Light pink shading" indicates a greater than 10% difference in the percentage of students tested.

▨ "Light peach shading" indicates differences in actual and estimated percent proficient that exceed 5 percent.

Estimating Proficiency Rates

Because each state implements its own tests and sets its own cut scores, we can't directly compare a Wisconsin test result to one in North Dakota. Several previous studies, however, have made comparisons among state tests by aligning their cut scores to a common instrument. Most of these aligned proficiency cut scores to the scale used for the National Assessment for Educational Progress (McGlaughlin et al. 2008; NCES 2007; Qian and Braun 2005; McGlaughlin and Bandeira de Mello 2002, 2003; McGlaughlin 1998a,1998b). NWEA's MAPs were used to estimate state cut scores for *The Proficiency Illusion* and other studies (Cronin et al. 2007a; Kingsbury et al. 2003). Results on the MAP assessment were combined with the estimated cut scores for this test to estimate proficiency rates for the sample.

MAP tests are computer-adaptive assessments in the basic skills. Starting in grade two and continuing through high school, these tests are taken by students in more than 3,000 school systems in 49 states and several foreign countries. The MAP tests were developed in accordance with the test design and development principles outlined in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999). The *Computer-Based Testing Guidelines* (2000) of the Association of Test Publishers and the *Guidelines for Computerized-Adaptive Test Development and Use in Education* (American Council on Education 1995) are used to guide test development and practices related to NWEA's use of computer-adaptive testing. Content on the MAP assessments is aligned to the curriculum standards for each state in which it is used, so that the test is a reasonable reflection of the content that students are expected to learn in each state. Because evidence related to the general content validity of MAP assessments is available in Appendix 1 of *The Proficiency Illusion*, we refer interested readers to that document for a more complete discussion of the assessment, its measurement characteristics, and the associated scale.

To estimate proficiency cut scores for *The Proficiency Illusion*, we created a sample population of students who took both MAP tests and their respective state assessment. Next we calculated the proportion of students in this sample population who performed at a proficient or above level on the state test. Once this was known, we found the score on the MAP scale that would produce an equivalent proportion of students. For example, assume that students must achieve a score of 300 on their state test and that 75% of our sample population achieved that score. If 75% of that sample performed at a scale score of 200 on the MAP assessment, a score of 200 on the MAP score would be equivalent to the state passing score of 300. This is a common method for estimating cut scores across tests and is known as the equipercentile or distributional method.

To evaluate the efficacy of this method, a pilot study of five states was conducted in which the distributional method was used to evaluate how accurately cut scores from one sample predicted the proficiency status of individual students in a second sample in each state (Cronin, Kingsbury, Bowe, & Adkins, 2007b). The results indicated that the cut scores estimated from MAP testing with the first sample accurately predicted the proficiency status of 84% of the students in the second sample in reading and 86% of the students in math. In addition, when applied to the entire sample, the predicted proficiency rate for the sample in each state fell within an average of 3 percentage points of the actual results for the group in reading, and within an average of 2.1 percentage points of the actual results in math.

The latter finding is particularly important for purposes of this study, because it demonstrated that when the estimated MAP cut scores are used, a school's projected proficiency results on the MAP assessment consistently came within 3 points of duplicating its actual results on its state assessment. This means that these methods for estimating cut scores can also be applied to make a reasonable prediction of a school's approximate proficiency rate on its state test.

The cut scores reported in *The Proficiency Illusion* were used for 25 of the states in the sample. These cut scores were estimated from data collected during the spring 2005, fall 2005, or spring 2006 testing terms. An additional 3 states were included in this study and data for these estimates came from spring 2007 testing data. Sampling data associated with the 25 states studied can be found in Appendix 3 of *The Proficiency Illusion*. The projected MAP percentile ranks associated with proficiency in the 28 states in this study are reported in Appendix B and Appendix C of this document.

The estimated cut scores for each of the states were applied to the 36 sample schools' spring 2006 MAP results in reading and math in order to determine the projected proficiency status of each student relative to each state's standard. Accordingly, students whose MAP scores were equal to or greater than the projected cut score for a state were identified as proficient in that state. From this information, we calculated an estimate of the overall proficiency rate that represented the proportion of students who scored proficient at each school, and derived an estimate of the proficiency rate for the subgroup populations within each school.

Estimating the AYP Status of Schools

The intent of NCLB is to ensure that 100% of each school's students achieve proficient performance in reading and math by the year 2014. To hold schools accountable for progress toward this goal, states set gradually escalating benchmark rates for proficiency that must be achieved by schools each year. These benchmarks, called AMOs, must not only be achieved by the student population as a whole, but also by ethnic subgroup members, low-income students, SWDs and LEP students whose group size exceeds the minimum count required by the state. NCLB also requires at least 95% of the school's enrolled students to take the standard version of the state test, and directs states to identify another indicator of school performance beyond test scores. States generally use attendance as the indicator for elementary and middle schools.

In order to make AYP, schools must meet all the criteria with each and every subgroup. Failing to make AYP for two consecutive years leads to the imposition of sanctions that escalate if the school fails to meet AYP in successive years. Sanctions range from requiring that schools offer students an opportunity to transfer after their school fails to make AYP for two consecutive years, to eventually closing or reconstituting the school after it fails for six years in a row.

For schools that do not meet the proficiency requirement for any subgroup, many states employ a confidence interval as a safety net. The confidence interval is a statistical measure that provides a margin of error, much like that reported as part of public opinion polls. If the observed proficiency rating for a failing subgroup, plus the estimated margin of error, meets the required proficiency rating, that subgroup is still considered to have met that AMO.

For example, assume that Washington Elementary School (a hypothetical school) tests 100 students from Subgroup E in reading, and assume that a 50% proficiency rate is required to meet the AMO for that group. But only 49 students (49%) pass the reading test. If a 95% confidence interval around the observed pass rate were applied, it might yield a margin of error of approximately ±4 points, depending on the variability within the sample. Consequently, the confidence interval around the observed pass rate would be 49% plus or minus 4 points, or 45% to 53%. Because the upper range of this interval (53%) exceeds the pass rate of 50% required to meet the AMO in this example, that subgroup would have passed.

Schools that fail to meet the proficiency testing requirements required by NCLB in any given year may also meet an AMO if they meet the criteria necessary to qualify for the act's safe harbor provision. To do this, a school must reduce the number of nonproficient students within a failing subgroup(s) by at least 10% relative to the previous year. If that is accomplished, the school will meet the AMO for that subgroup if at least one additional academic criterion is met. The additional academic criterion varies across states and school levels (e.g., elementary versus high school), but may include attendance rates, graduation rates, percentages of students performing above proficient, or other such indicators. In our study, only a single year's performance data were available at the subgroup level, so it wasn't possible for us to evaluate whether a school might have achieved safe harbor status.

The entire set of rules governing AYP is extraordinarily complex. In addition, based on the data available to us, it wasn't possible to estimate the actual status of the schools in our sample against all the rules. For purposes of this study, then, we limited our evaluation of AYP status to the following rules:

- We evaluated whether the overall performance of students, as estimated by spring 2006 results on the NWEA assessment, would have been sufficient to meet the AYP proficiency target that the state had set for the 2007–2008 academic year.
- For all ethnic subgroups with counts that exceeded the minimum subgroup size for evaluation, we determined whether their performance, as estimated on the spring 2006 NWEA assessment, was sufficient to meet the proficiency target set by the state for the same school year. We used ethnic identifiers supplied by the school to assign students to a subgroup. Because these identifiers are not always consistent across school systems, each student had to be reclassified into one of five ethnic subgroups: White, African American, Hispanic/Latino, Asian/Pacific Islander, or American Indian/Alaska Native. Students who were identified as mixed-race, such as White and Native American, were classified with the respective nonwhite subgroup. Students of unknown or unspecified race were removed from the analysis.
- All SWDs in a given school were included in the school's sample if they also took some form of their state's assessment. If the count for this subgroup exceeded the minimum subgroup size for evaluation, we determined whether its performance met the AMO for this subgroup.
- All LEP students in a given school were included in the school's sample if they also took their state's assessment. Once again, they were evaluated against the AMO if the count exceeded the minimum size.
- All low-income students in a given school were included in the sample if they also took their state's assessment. This subgroup was evaluated against the AMO when its count exceeded the minimum size.
- Students were evaluated in each subgroup for which they qualified. Consequently, the test result of an Asian student who had been classified as LEP would be considered three times, once when determining whether the school as a whole met its AMO, once when considering whether the Asian/Pacific Islander subgroup met its AMO, and once when considering whether the LEP group met that AMO. This application is consistent with the current NCLB rules (Sunderman 2006).
- For states that used confidence intervals as part of their AYP calculation, we applied the calculation in circumstances when a subgroup's performance fell short of meeting the required proficiency rate. Some states apply confidence intervals to the proficiency rate; others apply confidence intervals to student scores. Some use two-tailed tests; others use one-tailed. In each case, we applied the method the state reported using for calculating the confidence interval.

States have some leeway to make changes in their plans, subject to approval by the U.S. Department of Education. These changes may include the setting the trajectory for proficiency improvement rates, defining minimum subgroup sizes, and employing confidence intervals. We used the state accountability plans that were in place as of February 2008 (U.S. Department of Education 2008) as the primary form of documentation and applied the rules in place at that time to conduct the analysis.

Because schools report much of the information about subgroups to NWEA separately from their reports to the state, the subgroup identifiers supplied to us for this study are not always identical to those supplied to the state, particularly in terms of student ethnicity. This is one reason we caution that this study does not attempt a formal replication of any particular school's state test results and AYP status. Instead, we aim to illustrate how a school with the particular data supplied to us might perform relative to some of the various states' standards and AYP rules.

For this analysis, then, we attempted to determine the

AYP status of a fixed group of students at a single point in time against the AYP targets for 2008. We included all subgroups that exceeded the minimum size in the analysis and applied confidence intervals for those states in which it was appropriate. We didn't evaluate safe harbor status, participation rates in state testing, growth models, or average daily attendance in this study, nor did we attempt to evaluate whether a school had met NCLB requirements for bringing adequate numbers of highly qualified teachers on board.

APPENDIX B

Table B-1. Estimated state test proficiency cut scores in reading using MAP*

State	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade
Arizona	23	25	25	32	30	36
California	61	43	53	56	52	56
Colorado	7	11	11	13	17	14
Delaware	28	32	23	27	23	20
Florida	33	40	53	34	37	50
Georgia	16	16	12	7	12	8
Idaho	33	32	32	34	37	36
Illinois	35	27	32	25	32	22
Indiana	27	27	29	32	34	33
Kansas	35	29	40	32	32	33
Maine	37	43	44	46	43	44
Massachusetts	55	65	50	43	46	31
Michigan	16	20	23	21	25	28
Minnesota	26	34	32	37	43	44
Montana	26	25	27	30	32	36
Nevada	46	40	53	34	40	39
New Hampshire	33	34	34	43	40	48
New Jersey	15	25	16	27	23	n/a
New Mexico	33	32	30	43	32	33
North Dakota	22	29	34	37	30	33
Ohio	21	21	21	25	23	22
Rhode Island	33	34	34	43	40	48
South Carolina	43	58	64	62	69	71
Texas	12	23	30	21	32	n/a
Vermont	33	34	34	43	40	48
Washington	37	23	27	40	49	36
Wisconsin	14	16	16	16	17	14
Wyoming	49	49	44	52	43	44
28-state median	33	29	32	34	32	36

*In percentile ranks; n/a = not available

APPENDIX C

Table C-1. Estimated state test proficiency cut scores in math using MAP*

State	3rd grade	4th grade	5th grade	6th grade	7th grade	8th grade
Arizona	30	28	33	40	36	42
California	46	55	57	62	59	64
Colorado	6	8	9	16	19	25
Delaware	25	26	24	29	36	36
Florida	30	40	46	52	43	32
Georgia	8	23	10	33	22	15
Idaho	30	34	35	38	41	47
Illinois	20	15	20	20	19	20
Indiana	35	32	31	27	26	34
Kansas	30	34	35	33	45	38
Maine	43	46	46	52	54	53
Massachusetts	68	77	70	67	70	67
Michigan	6	13	21	27	35	32
Minnesota	30	43	54	52	52	51
Montana	43	43	40	45	43	60
Nevada	50	46	46	35	36	38
New Hampshire	41	35	34	44	44	53
New Jersey	13	23	26	40	43	n/a
New Mexico	46	49	54	60	61	56
North Dakota	20	27	23	32	39	41
Ohio	20	31	40	33	32	31
Rhode Island	41	35	34	44	44	53
South Carolina	71	64	72	65	68	75
Texas	30	34	24	35	41	n/a
Vermont	41	35	34	44	44	53
Washington	36	46	48	57	59	56
Wisconsin	29	29	26	21	21	23
Wyoming	36	43	43	42	45	51
28-state median	32.5	34.5	34.5	40	43	42

*In percentile ranks; n/a = not available

REFERENCES

American Council on Education. 1995. *Guidelines for Computerized-Adaptive Test Development and Use in Education*. Washington, DC: American Council on Education.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). 1999. *Standards for Educational and Psychological Testing*. Washington, DC: AERA, APA, and NCME.

Associated Press April 17, 2006. With help of states, U.S. government, schools duck potential penalties. http://www.msnbc.msn.com/id/12357165/from/RSS/ (accessed September 22, 2008).

Association of Test Publishers (2000). *Guidelines for Computer-Based Testing*. Washington, D.C.. Association of Test Publishers.

Chudowsky, N., and V. Chudowsky. 2008. Many states have chosen a back-loaded approach to No Child Left Behind goal of all students scoring proficient. Washington, DC: Center on Education Policy. http://www.cep-dc.org/index.cfm?fuseaction=document_ext.showDocumentByID&nodeID=1&DocumentID=238 (accessed September 19, 2008).

Council of Chief State School Officers (CCSSO). 2008. Profiles of state accountability systems, California state profile 2006–2007. http://accountability.ccsso.org/index.asp (accessed August 1, 2008).

Cronin, J., M. Dahlin, D. Adkins, and G.G. Kingsbury. 2007a. *The Proficiency Illusion*. Washington, DC: Thomas B. Fordham Institute.

Cronin, J., G.G. Kingsbury, M. Dahlin, D. Adkins, and B. Bowe. 2007b. Alternate methodologies for estimating state standards on a widely used computer-adaptive test. Paper presented at the Annual Conference of the American Educational Research Association, Chicago, IL.

Erpenbach, W.J., and E. Forte. 2005. *Statewide Educational Accountability under the No Child Left Behind Act—A Report on 2005 Amendments to State Plans*. Washington, DC: CCSSO.

Fulton, M. 2006. *State Note. Minimum Subgroup Size for Adequate Yearly Progress: State Trends and Highlights*. Denver, CO: Education Commission of the States. http://www.ecs.org/clearinghouse/71/71/7171.pdf (accessed September 22, 2008).

Kane, T. J., and D.O. Staiger. 2002. Volatility in school test scores: Implications for test based accountability systems. Pages 235–238 in *Brookings Papers on Education Policy*, edited by D. Ravitch. Washington, DC: Brookings Institution.

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

Kingsbury, G.G., A. Olson, J. Cronin, C. Hauser, and R. Houser. 2003. *The State of State Standards*. Lake Oswego, OR: Northwest Evaluation Association (NWEA).

Linn, R., and C. Haug 2002. *Stability of School Building Accountability Scores and Gains*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

McGlaughlin, D. H. 1998a. *Study of the Linkages of 1996 NAEP and State Mathematics Assessments in Four States*. Washington, DC: National Center for Education Statistics (NCES).

McGlaughlin, D. H. 1998b. Linking state assessments of NAEP: A study of the 1996 mathematics assessment. Paper presented at the American Educational Research Association, San Diego, CA.

McGlaughlin, D. H., and V. Bandeira de Mello. 2002. Comparison of state elementary school mathematics achievement standards using NAEP 2000. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

———. 2003. Comparing state reading and math performance standards using NAEP. Paper presented at the National Conference on Large-Scale Assessment, San Antonio, TX.

McLaughlin, D.H., V. Bandeira de Mello, C. Blankenship, K. Chaney, P. Esra, H. Hikawa, D. Rojas, P. William, and M. Wolman. 2008. *Comparison between NAEP and State Mathematics Assessment Results: 2003*. NCES 2008-475. Washington, DC: NCES, Institute of Education Sciences, U.S. Department of Education.

NCES. 2007. *Mapping 2005 State Proficiency Standards onto the NAEP Scales*. NCES 2007-482. Washington, DC: U.S. Department of Education.

National Education Association. 2006. NCLB testing results offer “complex, muddled” picture. http://www.nea.org/esea/ayptrends1104.html (accessed October 6, 2008).

Novak, J., and B. Fuller. 2003. Penalizing diverse schools? Similar test scores, but different students bring federal sanctions. Policy Brief. Berkeley: Policy Analysis for California Education (PACE).

NWEA. 2005. *Rit Scale Norms*. Lake Oswego, OR: NWEA.

Peterson, P., and F. Hess. 2008. Few states set world class standards. *Education Next* 8:3. http://www.hoover.org/publications/ednext/18845034.html (accessed September 19, 2008).

Porter, A., R. Linn, R., & and C.S. Trimble. 2005., C.S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice* 24(4): 32–39.

Qian, J., and H. Braun. 2005. *Mapping State Performance Standards on the NAEP Scale*. Princeton, NJ: Educational Testing Service.

Rogosa, D.R. 2003. The NCLB "99% confidence" scam: Utah-style calculations. http://www-stat.stanford.edu/~rag/nclb/utahNCLB.pdf (accessed October 3, 2008).

———. 2005. Statistical misunderstandings of the properties of school scores and school accountability. Pages 147 – 174 in *Yearbook of the National Society for the Study of Education*, edited by J. L. Herman and E. H. Haertel. Chicago, IL: National Society for the Study of Education.

San Francisco Chronicle, September 5, 2008. State falling way behind No Child Left Behind. http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2008/09/05/MNEJ12O85V.DTL&hw=Adequate+Yearly+Progress&sn=001&sc=1000 (accessed October 6, 2008).

Simpson, M.A., B. Gong, and S. Marion. 2005. *Effect of Minimum Cell Sizes and Confidence Intervals for Special Education Subgroups on School-Level AYP Determinations*. Dover, NH: National Center for Improvement of Educational Assessment.

Spellings, Margaret (2007, January). *Building on Results: A Blueprint for Strengthening the No Child Left Behind Act*. Washington, DC.: U.S. Department of Education.

Sunderman, G.L. 2006. *The Unraveling of No Child Left Behind: How Negotiated Changes Transform the Law*. Cambridge: The Civil Rights Project at Harvard University.

U.S. Department of Education. 2006. *National Assessment of Title I Interim Report: Executive Summary*. Washington, DC: Institute of Education Sciences.

———. 2008. *Approved State Accountability Plans. California State Plan.* http://www.ed.gov/admins/lead/account/stateplans03/index.html (accessed August 1, 2008).