THE Accountability Illusion

John Cronin, Michael Dahlin, Yun Xiang, and Donna McCahon

February 2009

Foreword by

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

THOMAS B. FORDHAM INSTITUTE
Advancing Educational Excellence

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

TABLE OF CONTENTS

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.

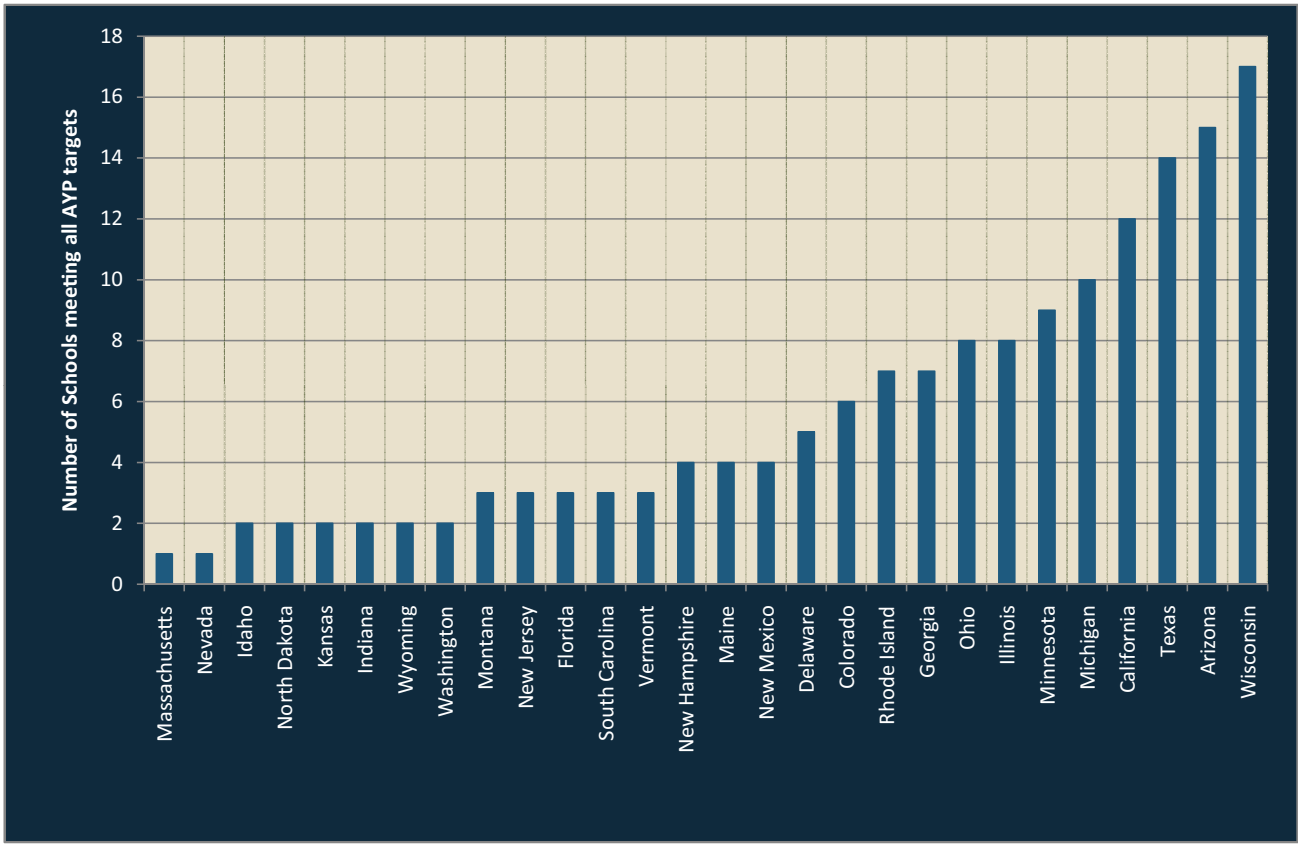

Figure ES-1. Number of sampled elementary schools that made AYP in 2008, by state

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.

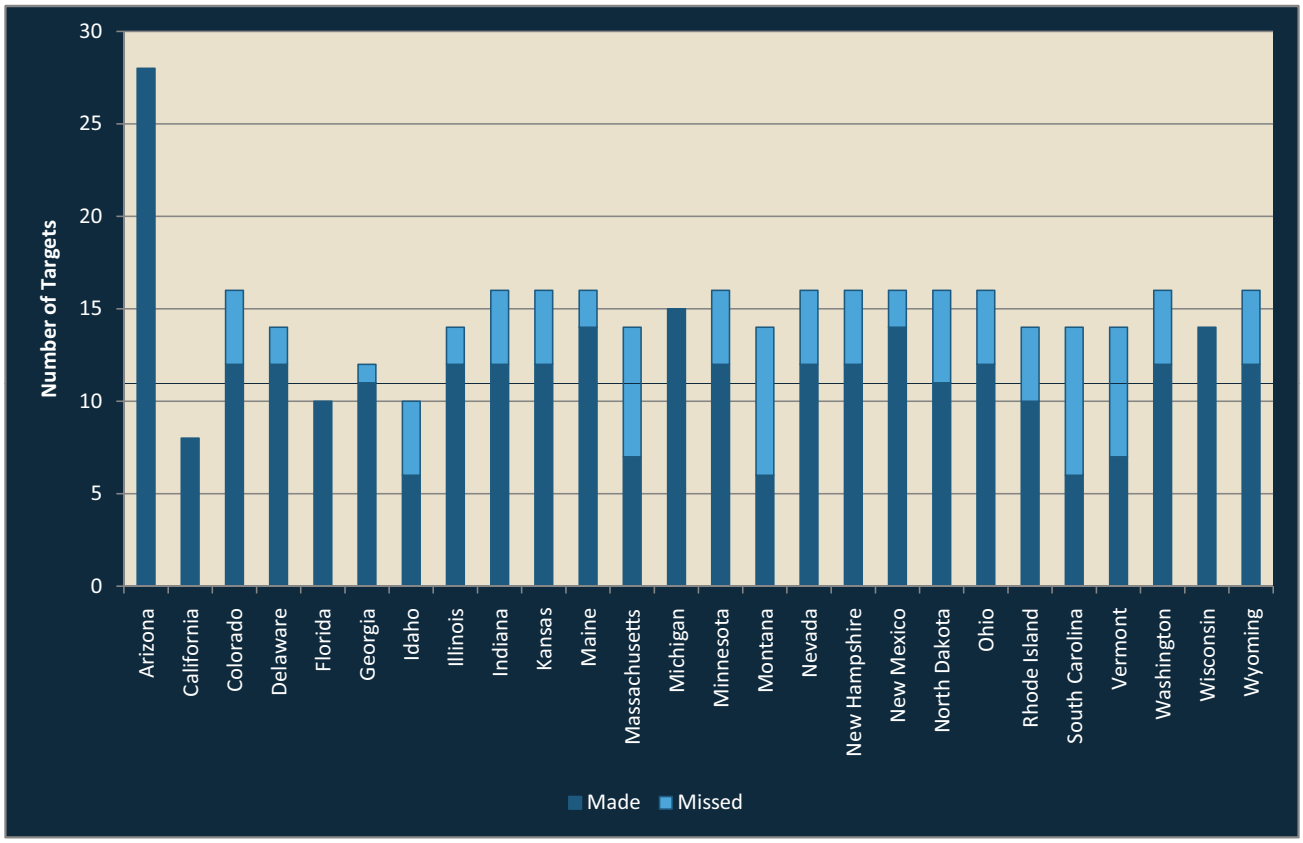

Figure ES-2. Number of subgroup targets met by Chaucer Middle School in 2008, by state

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.

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

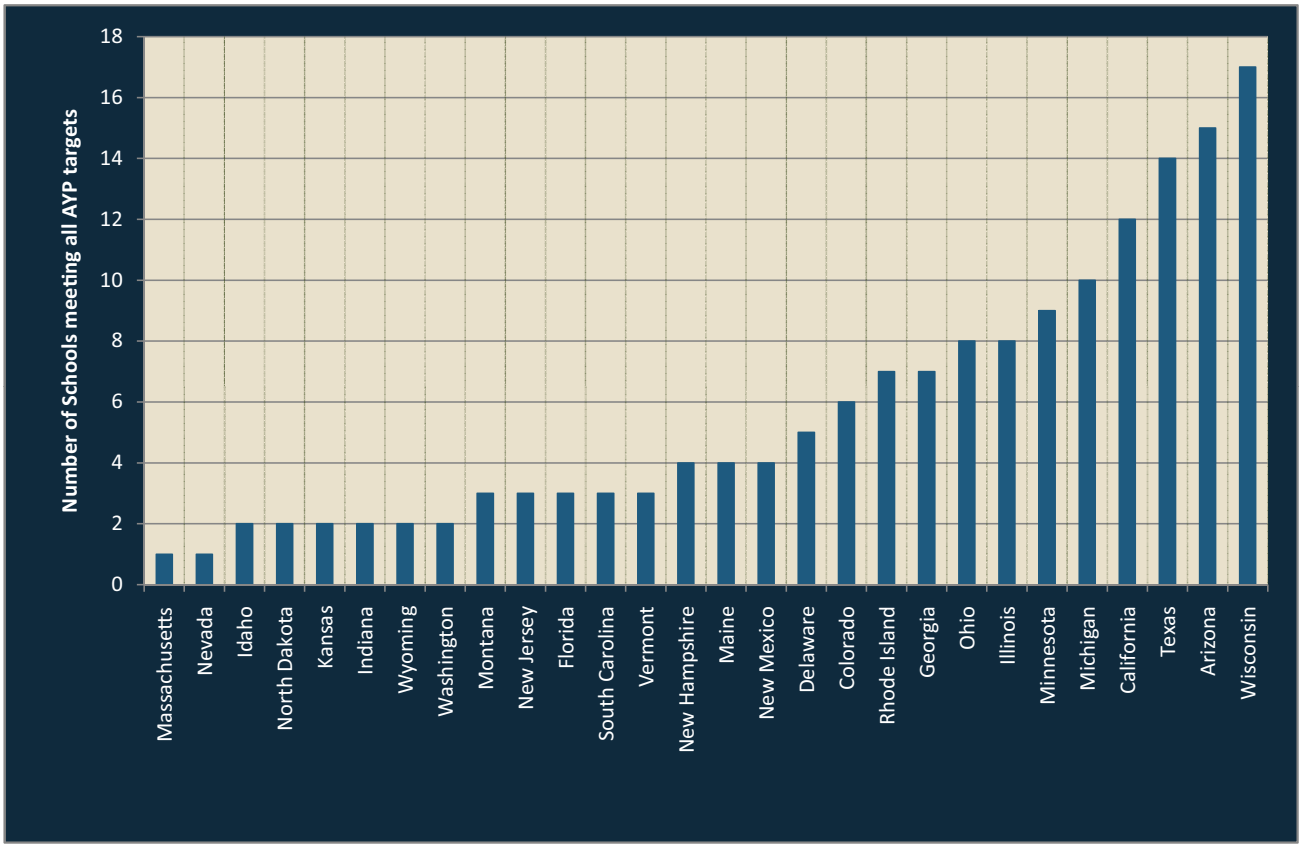

Figure 1. Number of schools in the elementary school sample making AYP by state (2008)

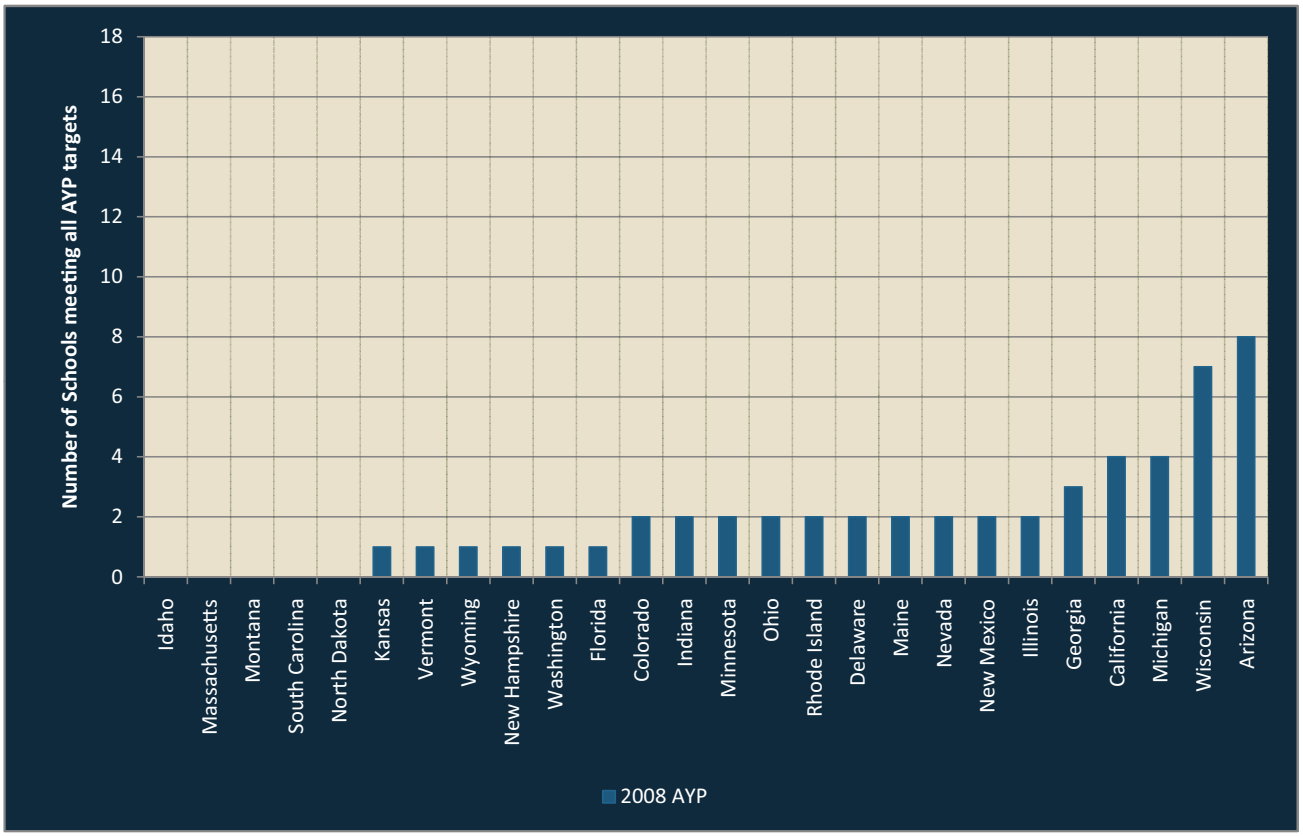

Figure 2. Number of schools in the middle school sample making AYP by state (2008)

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

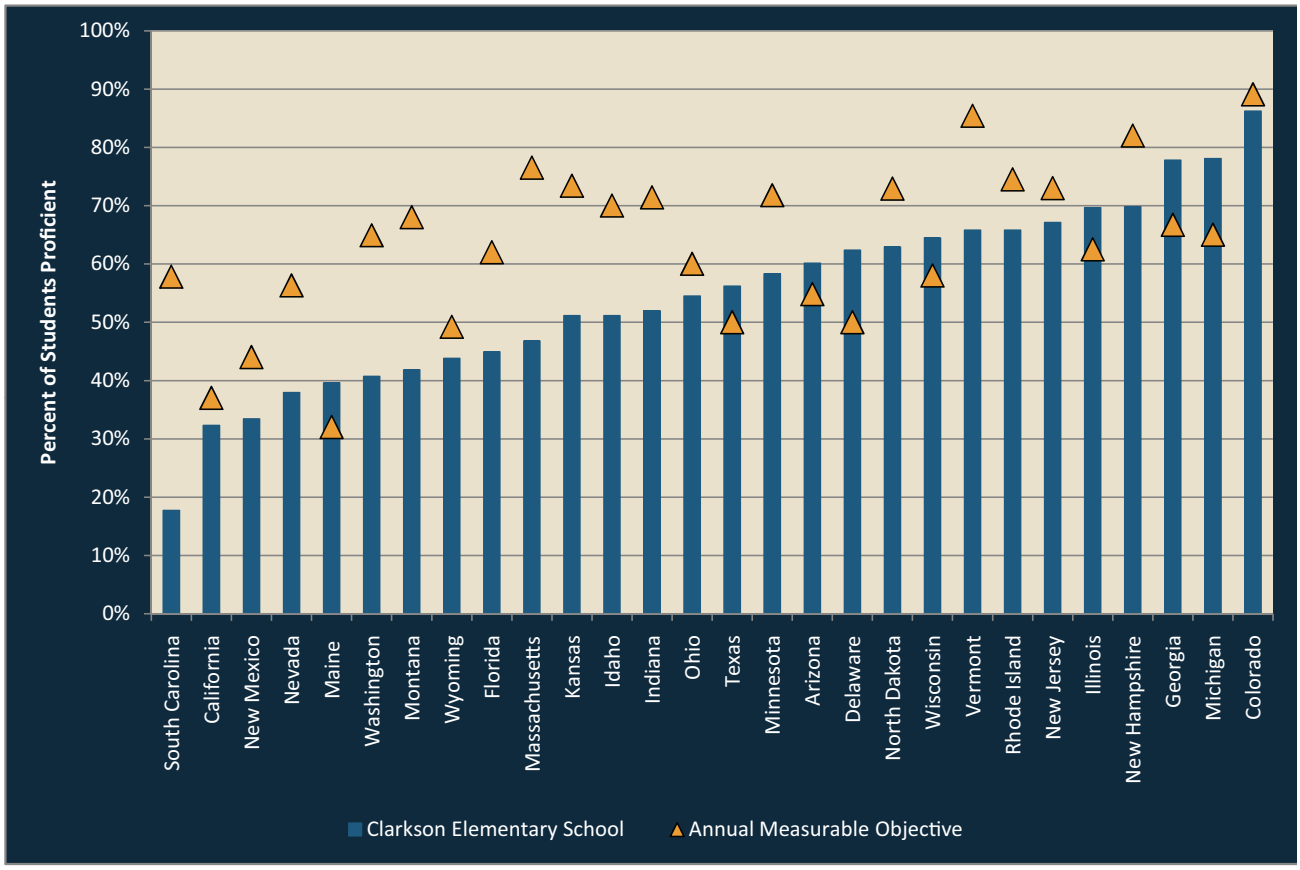

Figure 3. Math proficiency rate of Clarkson students relative to 2008 AMOs

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 assess-

ments. 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

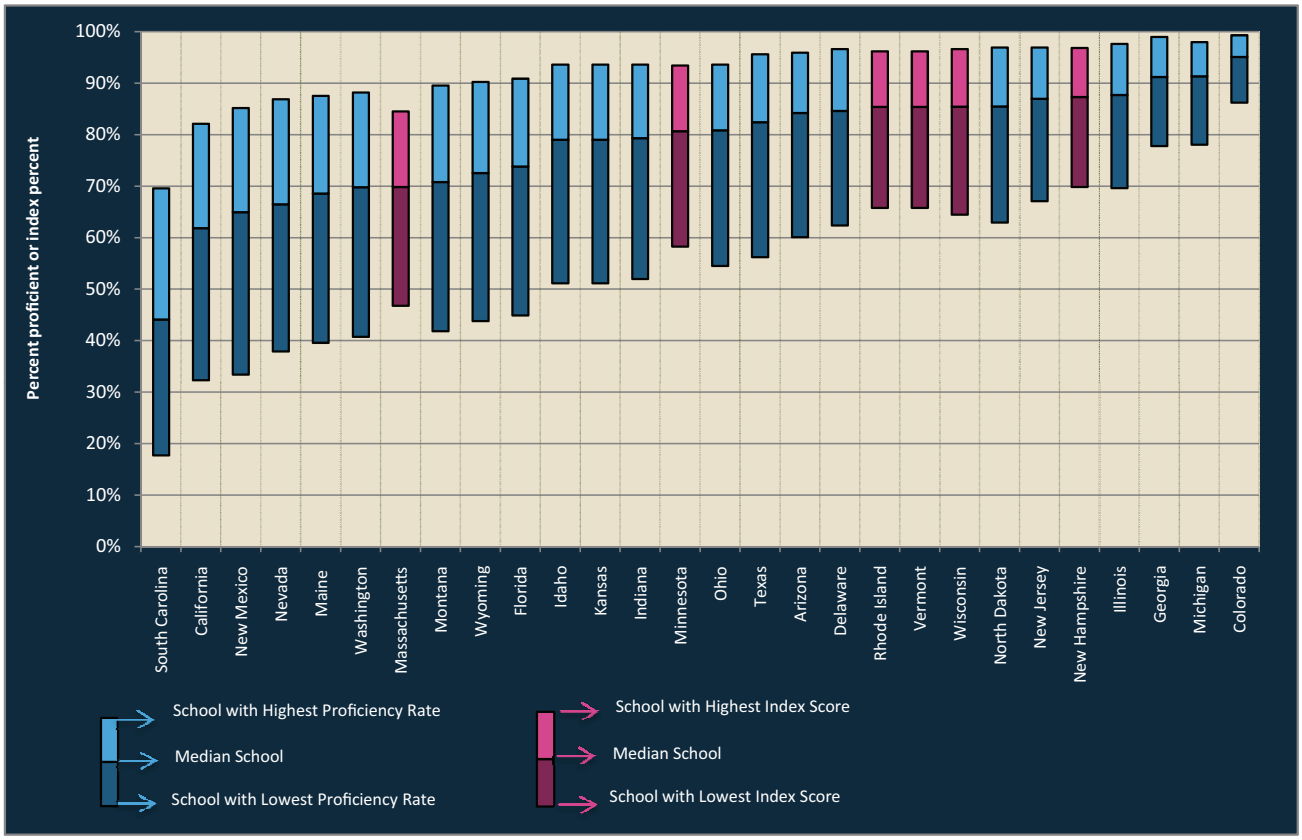

Figure 4. Overall proficiency rates of the elementary school sample in math

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.

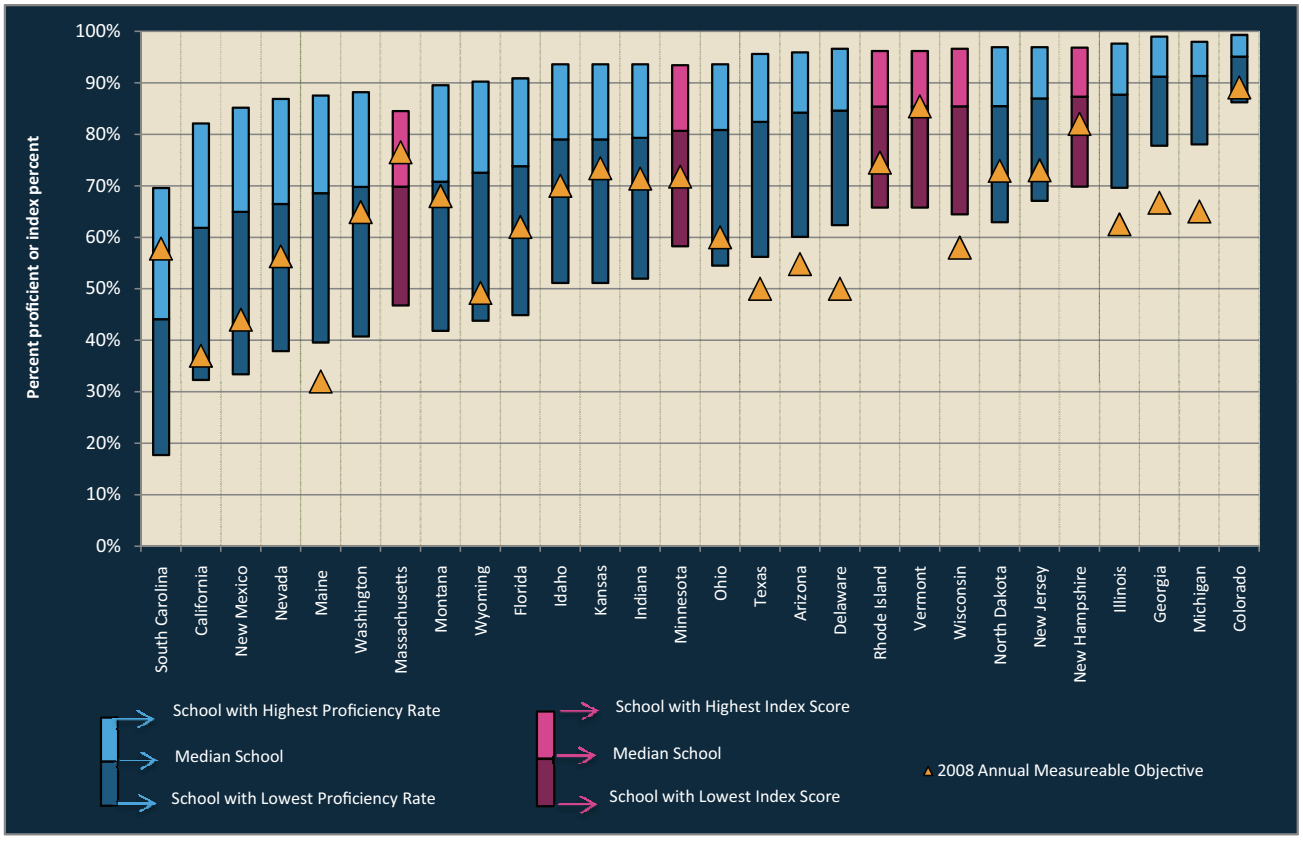

Figure 5. Math proficiency rates of the elementary school sample relative to each state's 2008 AMOs

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

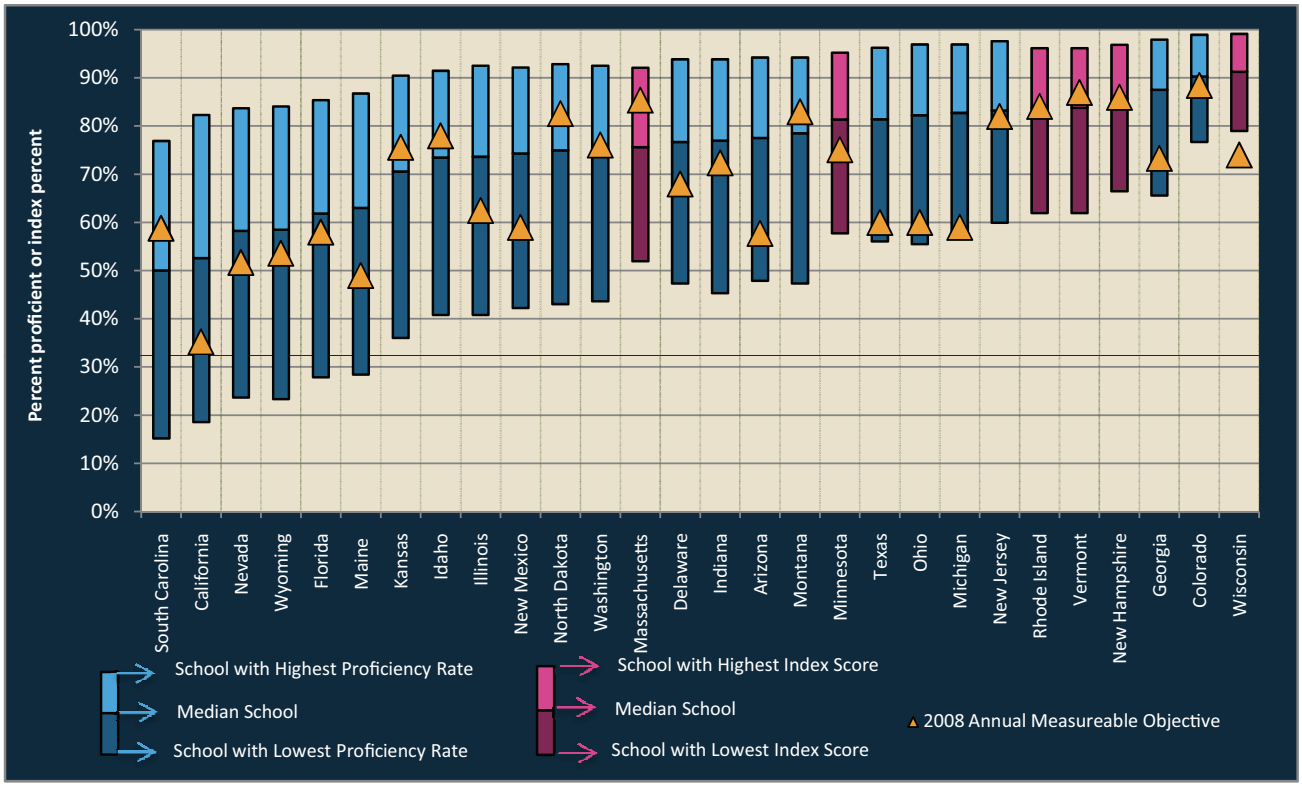

Figure 6. Reading proficiency rates of the elementary school sample relative to each state's 2008 AMOs

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-

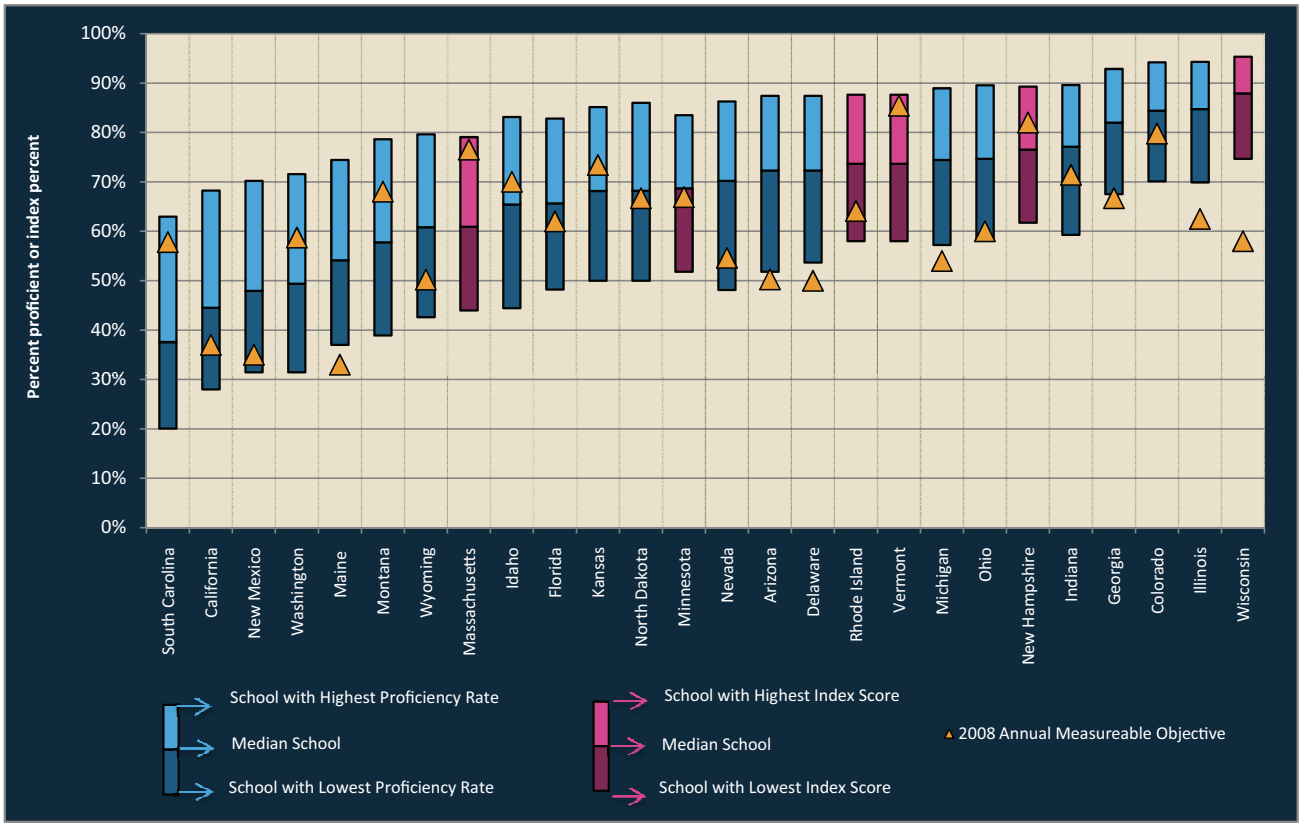

Figure 7. Math proficiency rates of the middle school sample relative to each state's 2008 AMOs

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.**

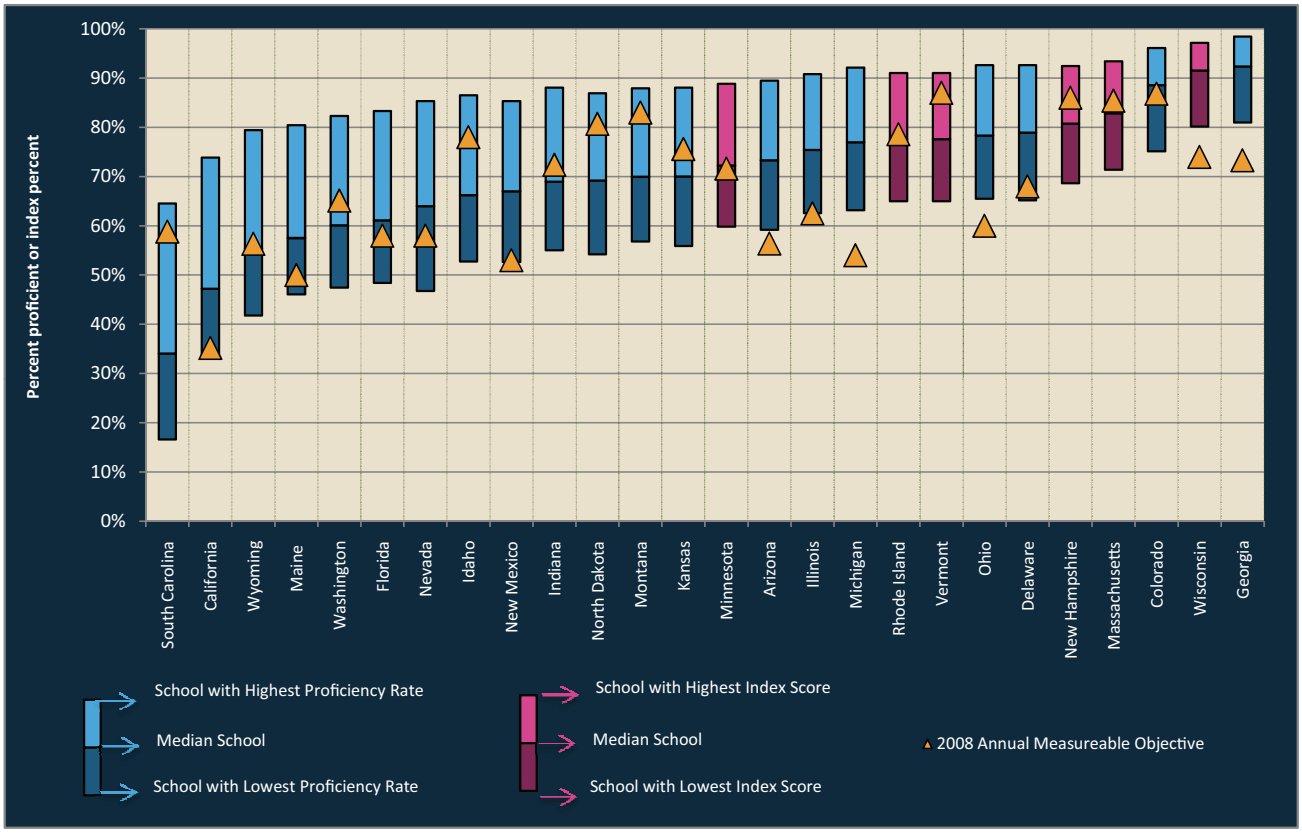

Figure 8. Reading proficiency rates of the middle school sample relative to each state's 2008 AMOs

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.

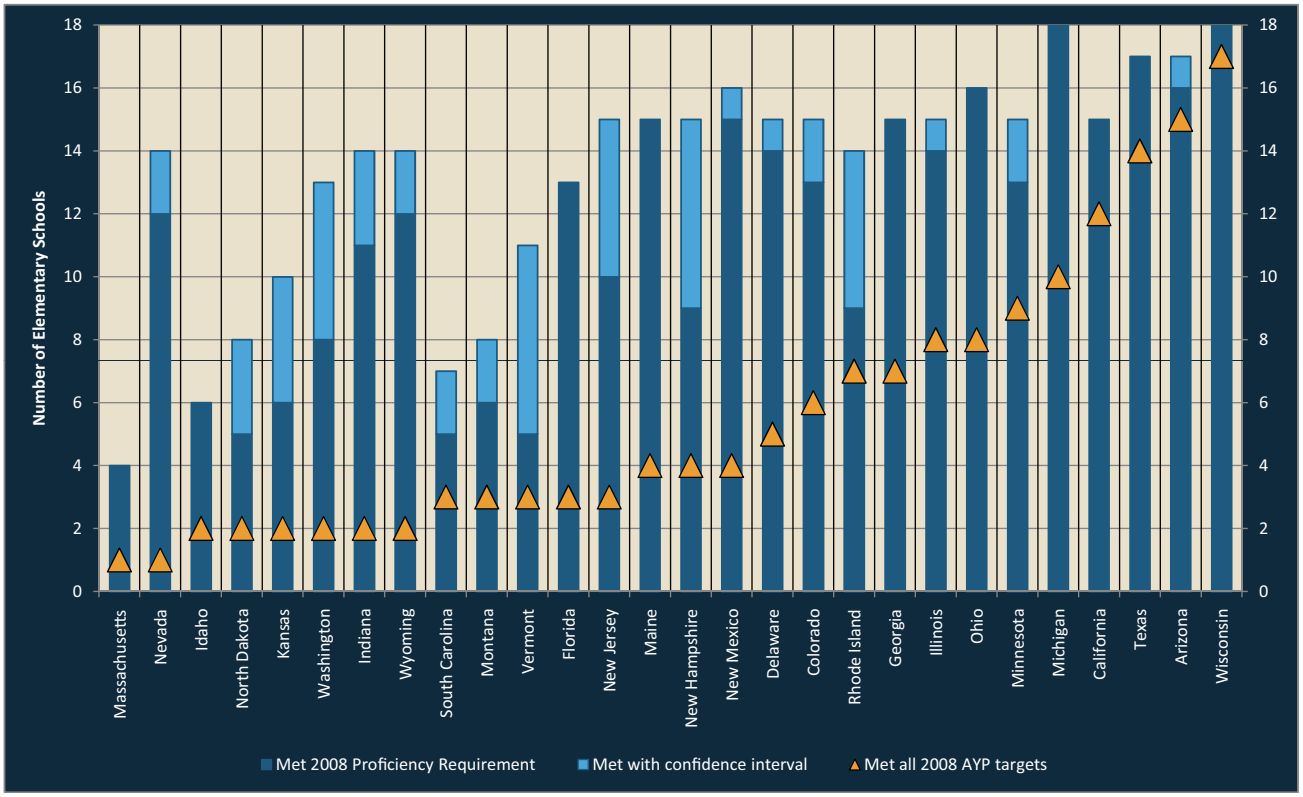

Figure 9. Number of elementary schools meeting 2008 AMOs with and without confidence intervals, by state

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.

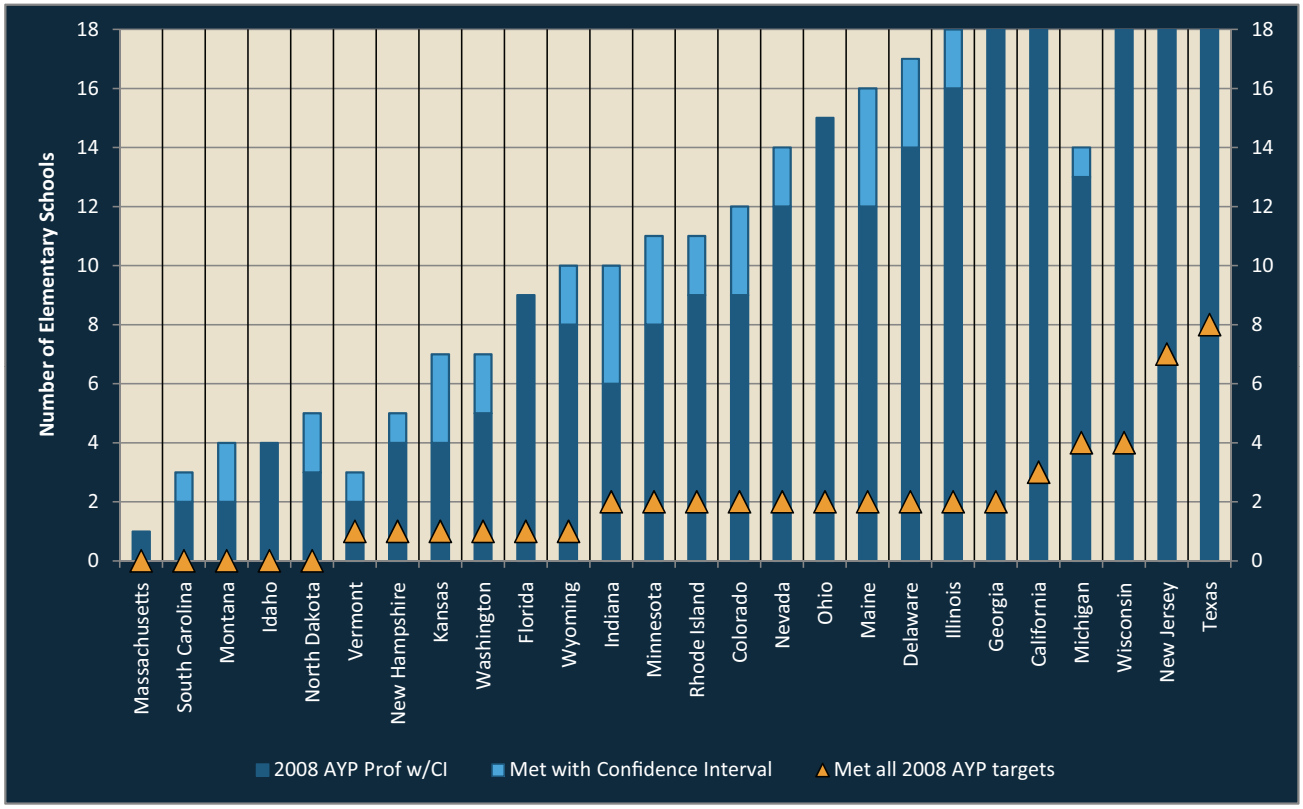

Figure 10. Number of middle schools meeting 2008 AMOs with and without confidence intervals, by state

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

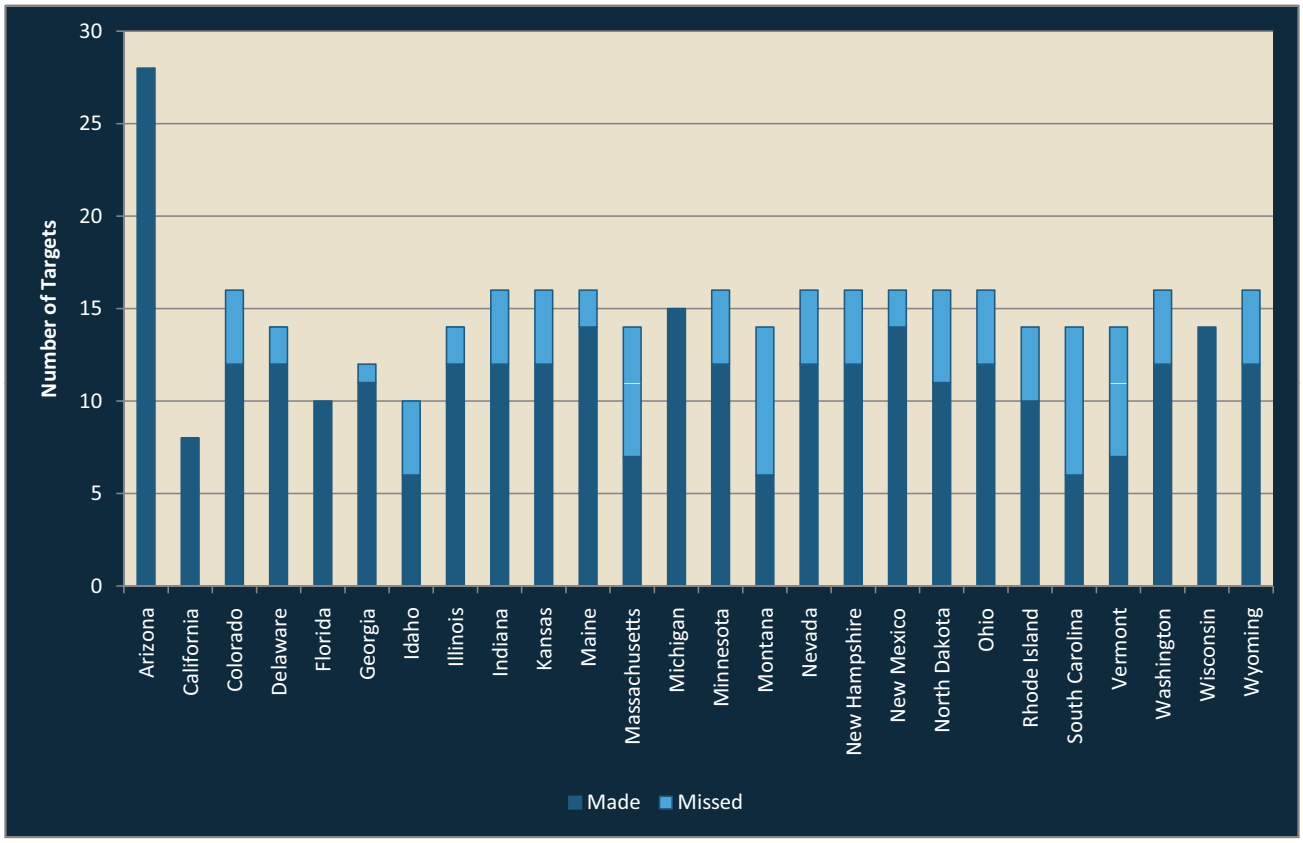

Figure 11. Number of subgroup targets met by Chaucer middle school in 2008

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.

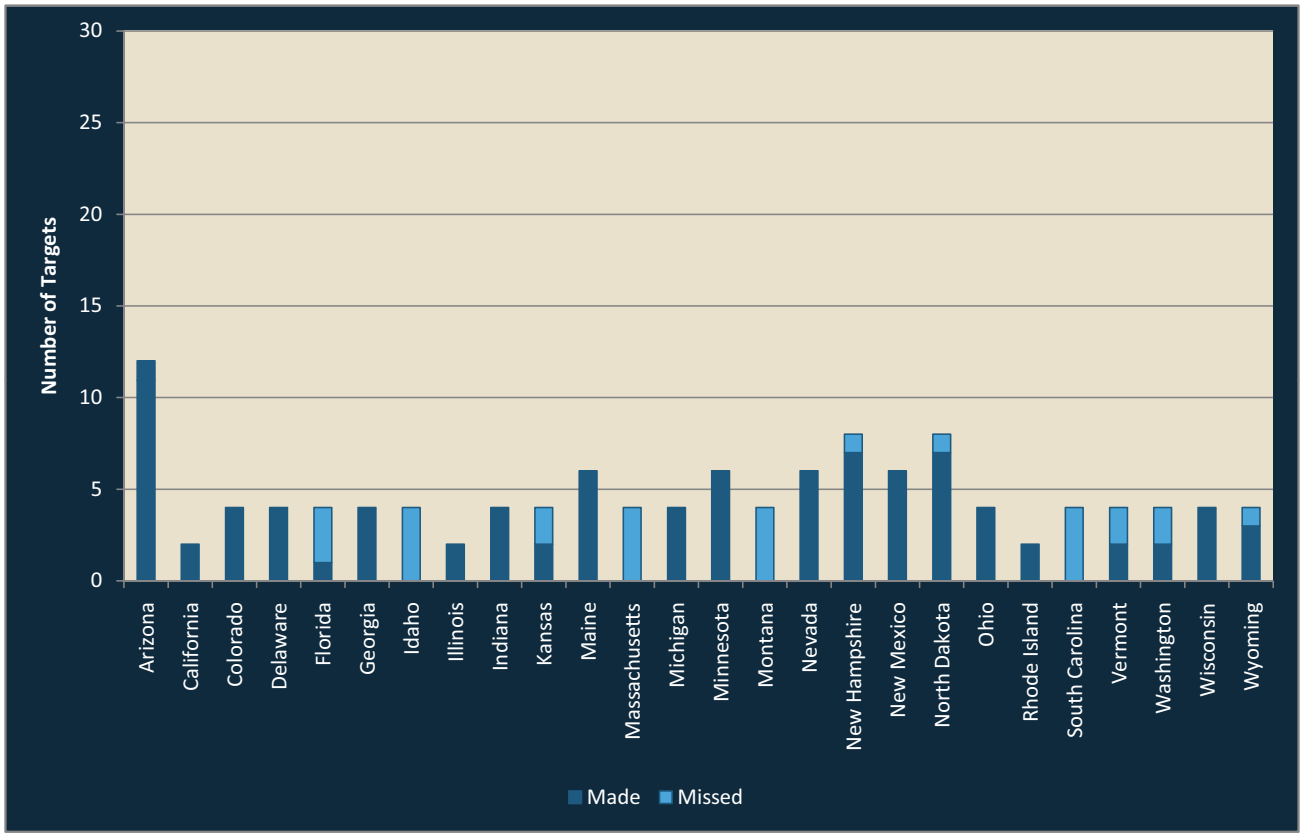

Figure 12. Number of subgroup targets met by Pogesto middle school (2008)

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%)

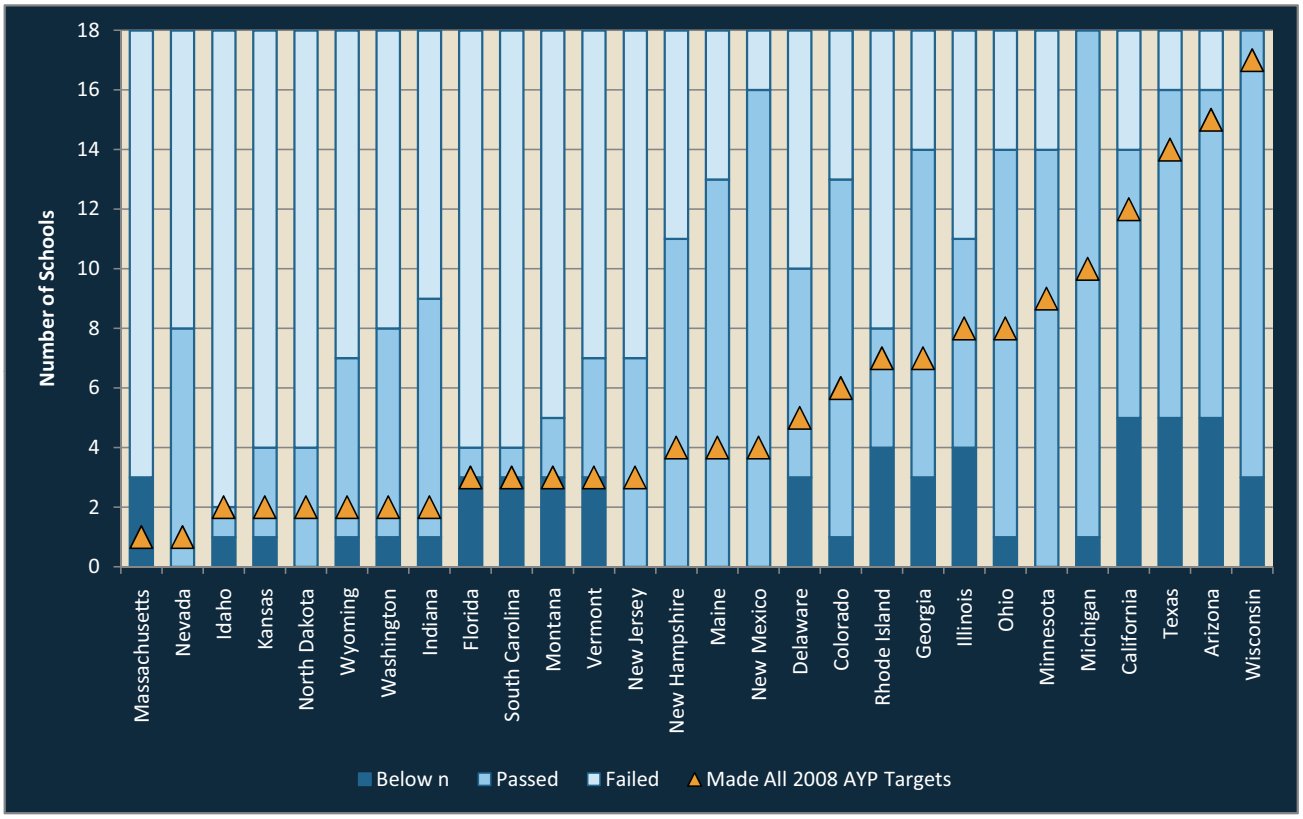

Figure 13. Number of elementary schools meeting 2008 AMOs in math and reading for their low-income student subgroup

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

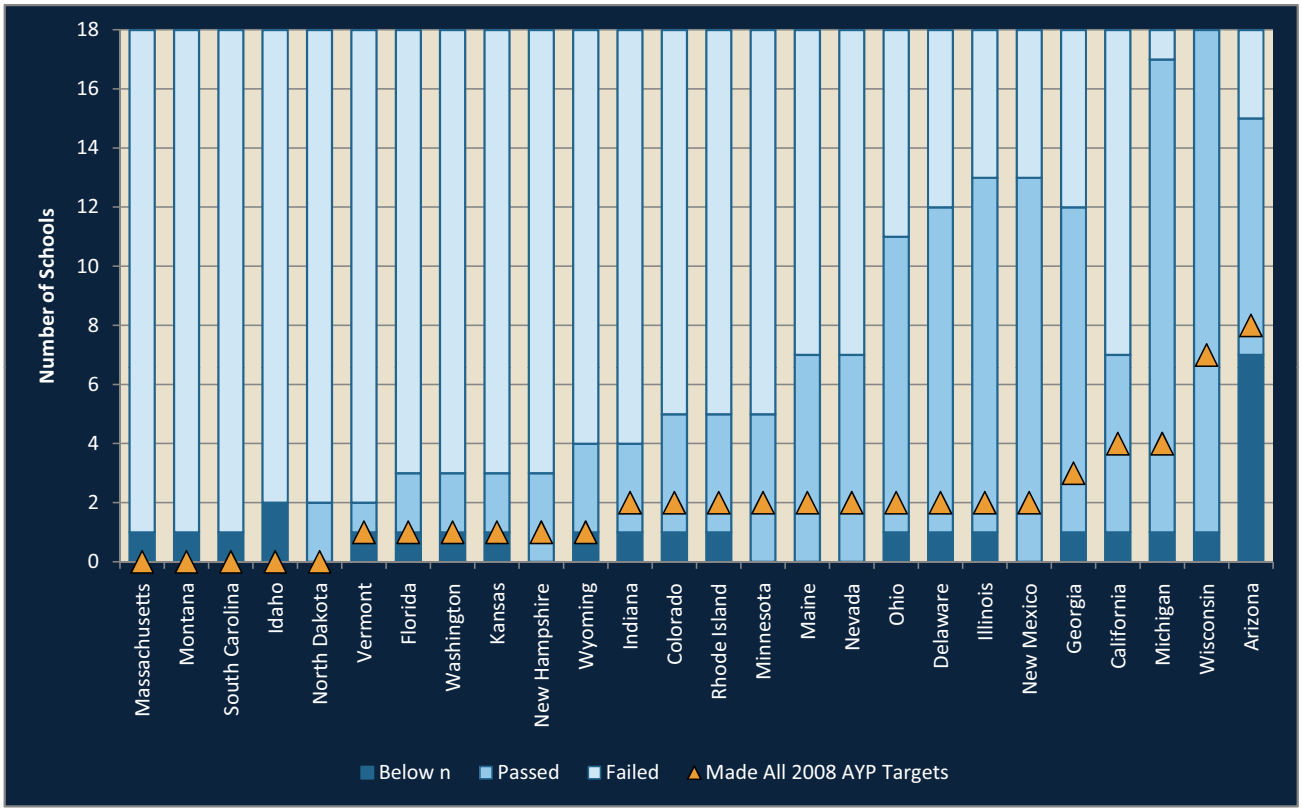

Figure 14. Number of middle schools meeting 2008 AMOs in math and reading for their low-income student subgroup

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.

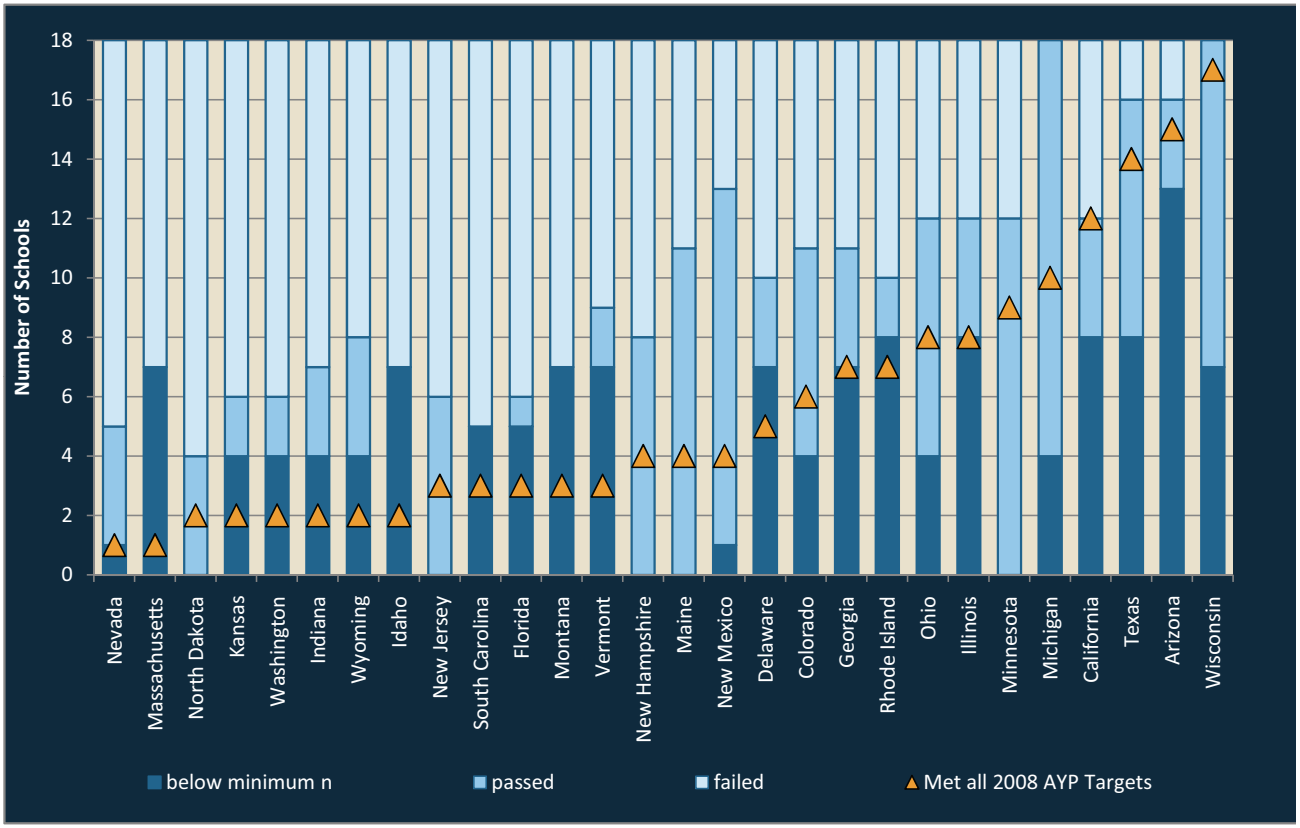

Figure 15. Number of elementary schools in which minority students met their 2008 AMOs

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.

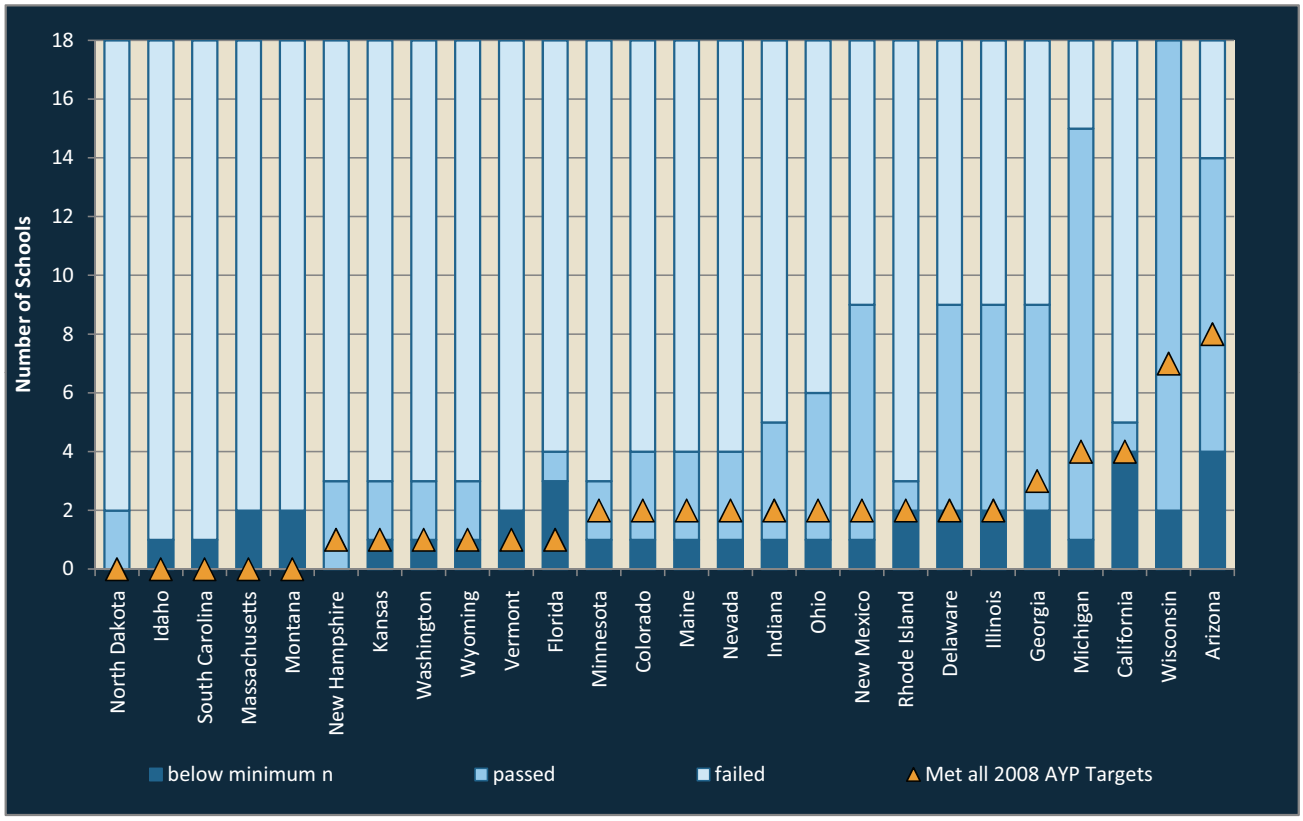

Figure 16. Number of middle schools in which minority students met their 2008 AMOs

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-

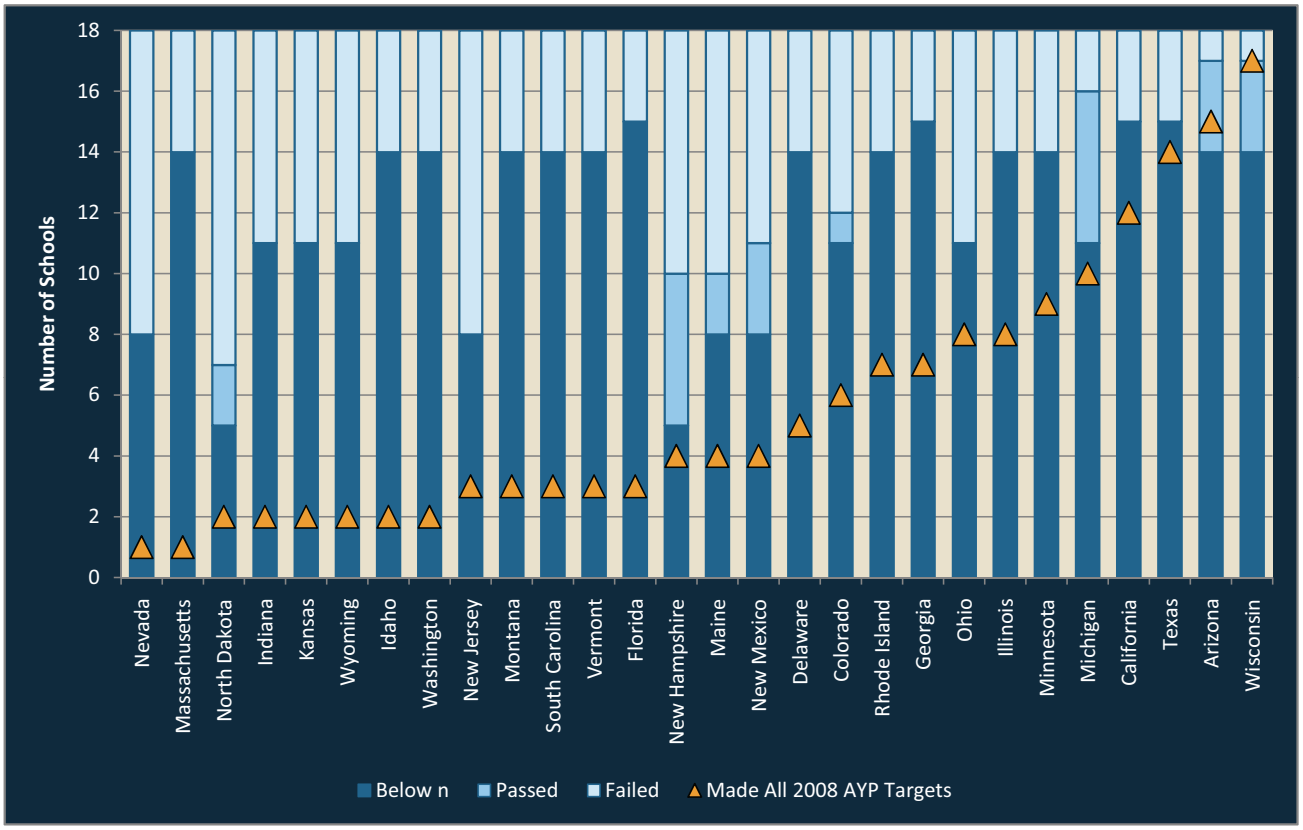

Figure 17. Number of elementary schools in which LEP students met their 2008 AMOs

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.

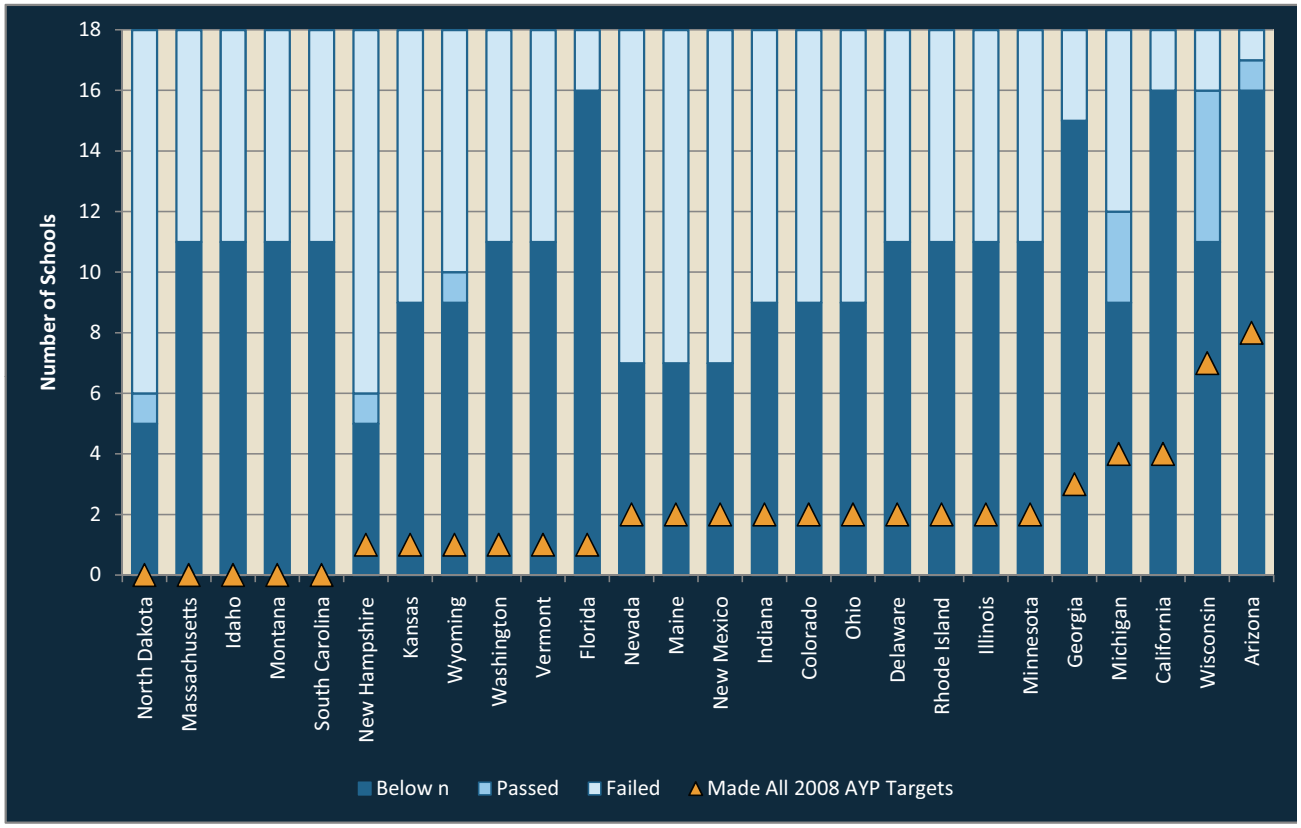

Figure 18. Number of sampled middle schools in which LEP students met their 2008 AMOs

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.

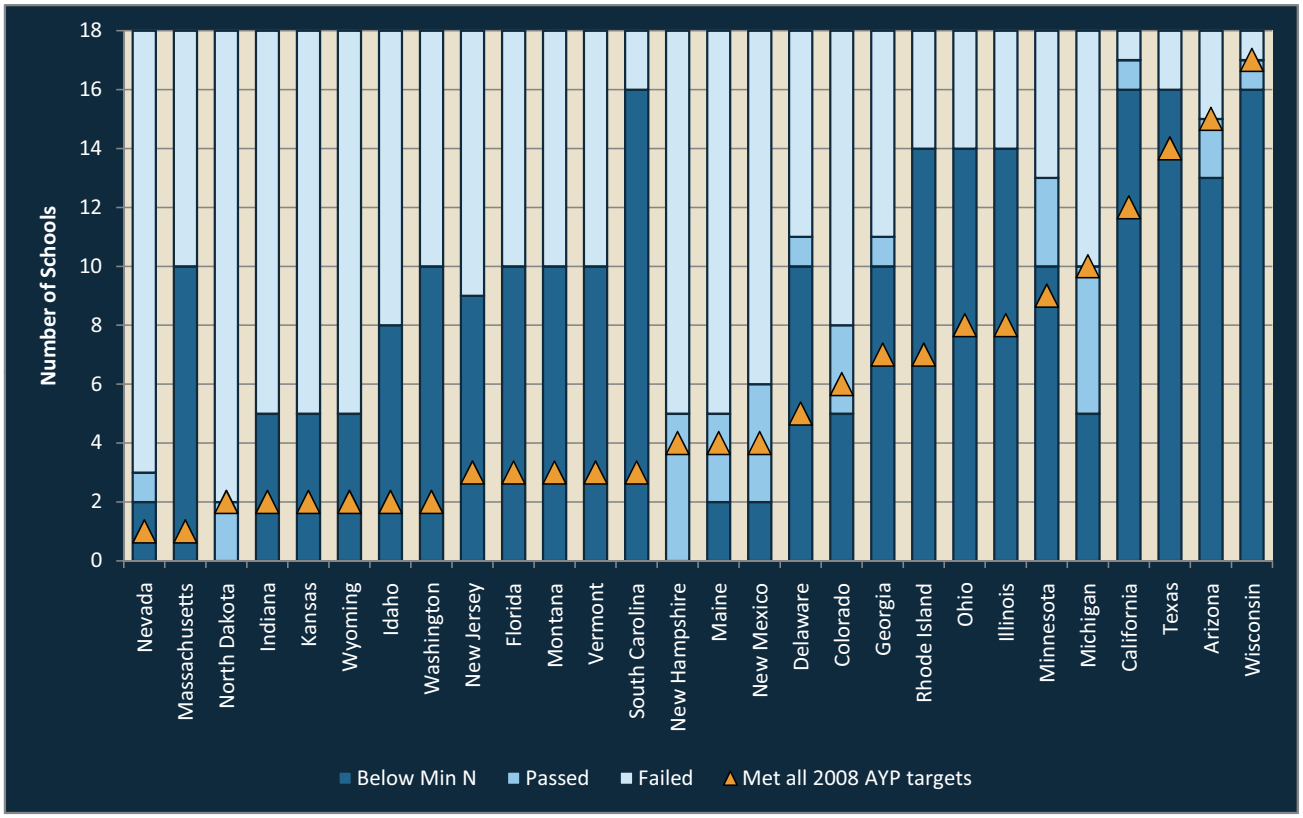

Figure 19. Number of sampled elementary schools in which SWDs met their 2008 AMOs

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.

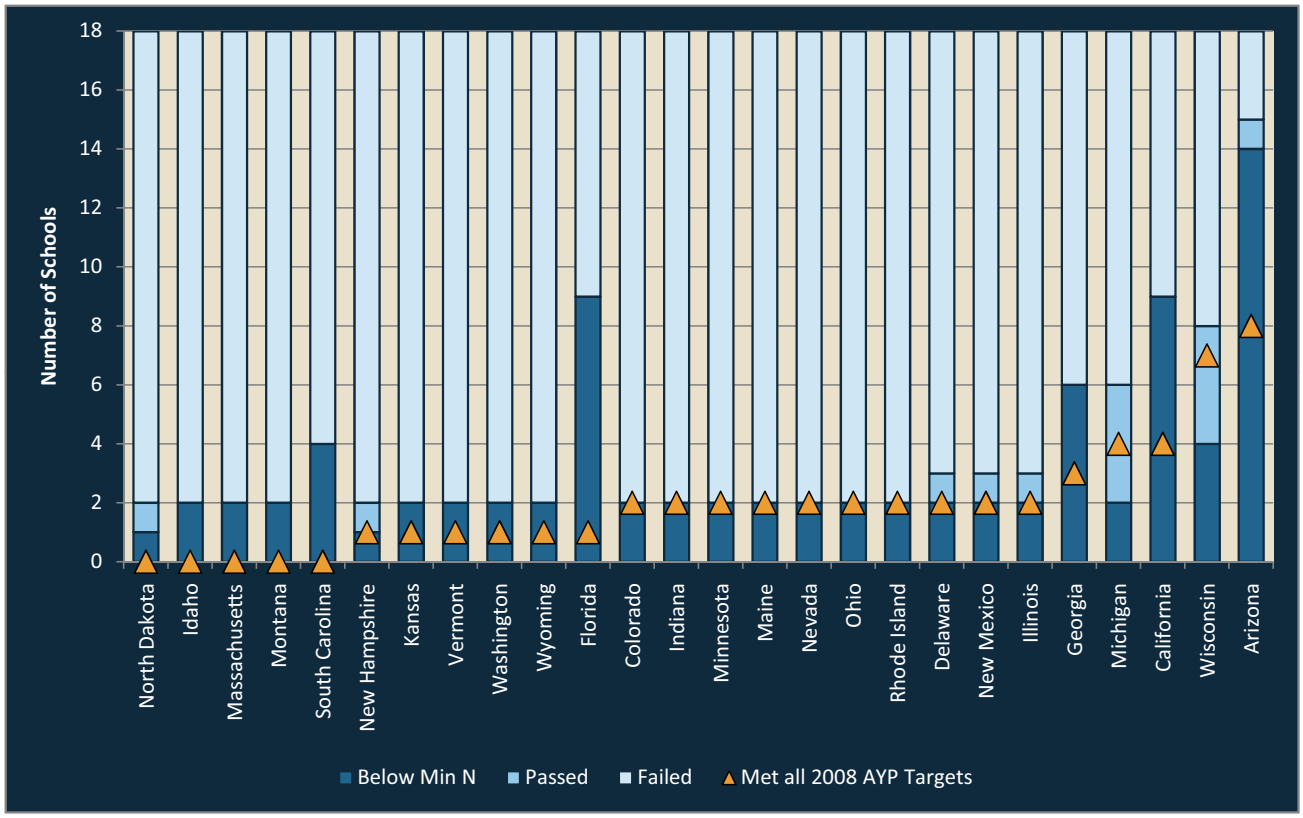

Figure 20. Number of sample middle schools in which SWDs met their 2008 AMOs

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

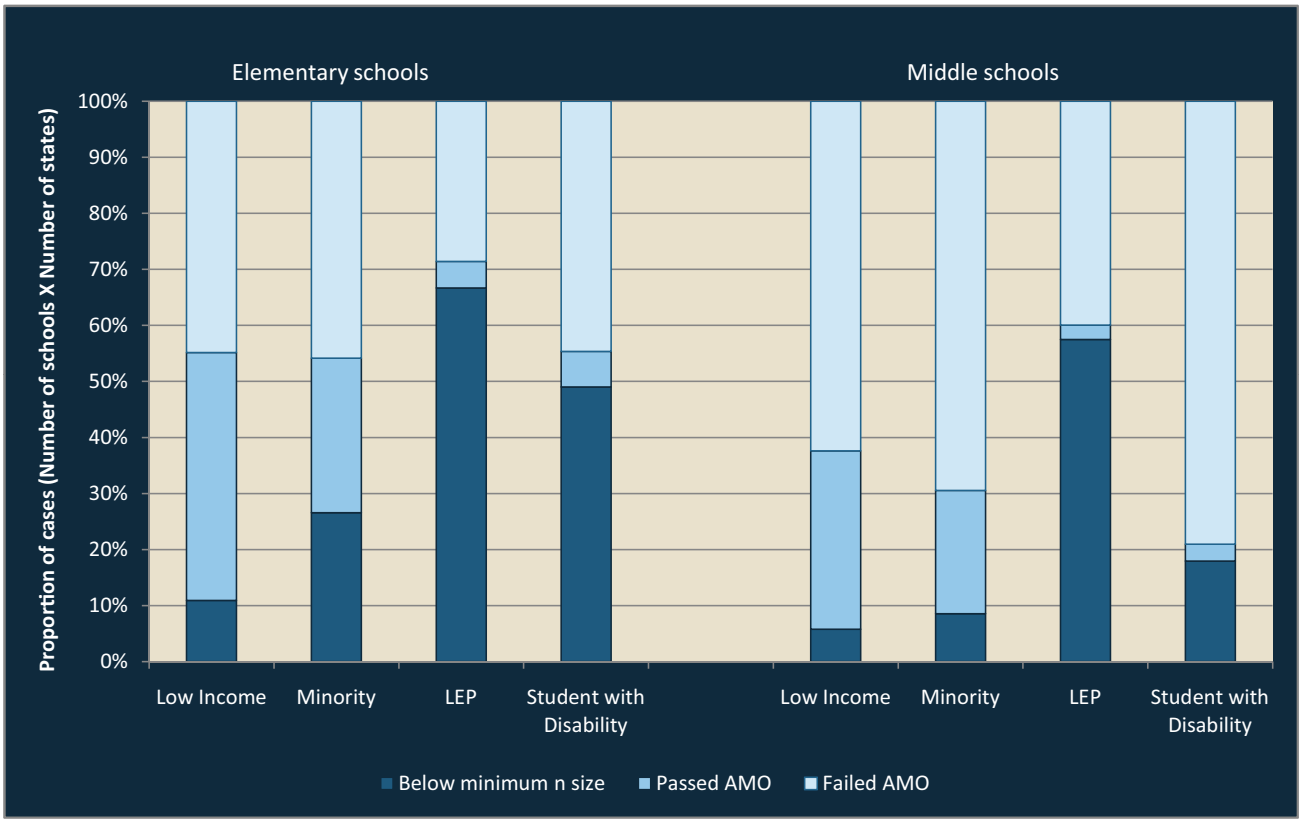

Figure 21. Summary of subgroup performance relative to AMOs

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 addi-

tional 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).

Arizona

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Arizona's NCLB accountability system—particularly how its various rules, criteria and practices result in schools either making AYP—or not making AYP. It also gauges how tough Arizona's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Arizona's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Arizona's AYP rules for the academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **3 of 18 elementary schools** and **10 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Arizona's accountability system. Among the 28 accountability systems examined in the study, there's only one state where more schools make AYP than in Arizona (Wisconsin). This makes The Grand Canyon State one of the least restrictive in terms of AYP passage rates (see Figure 1.)[2]
- Several sample schools made AYP in Arizona that failed to make AYP in most other states. This is probably because **Arizona's proficiency standards are relatively easy compared to other states** (**especially in reading**). Another reason is that Arizona's definitions for subgroups are grade-based rather than school based, resulting in fewer accountable subgroups (i.e., a school must have at least 40 individuals within a grade for that group to be evaluated). Arizona also uses a very generous confidence interval (or margin of error).

Arizona has several unique characteristics which contribute to the large number of schools making AYP in the state. In fact, only one other state in the study (Wisconsin) deems that more schools make AYP than Arizona does. One of the factors contributing to this is the rule set governing subgroup size. Unlike most states, Arizona considers each grade separately when determining whether a subgroup meets the criteria for accountability, which (for Arizona) is at least 40 students. For instance, a middle school in Arizona with three grades could have *almost* 120 African-American students, all performing poorly, and still make AYP as long as there are fewer than 40 African-American children in each grade. Another factor contributing to the high number of schools making AYP is Arizona's 99 percent confidence interval (i.e., statistical margin of error). This provides schools with greater leniency than the 95 percent confidence interval used by most other states in the study. Finally, Arizona's proficiency standards (or cut scores) are relatively easy in the early grades, compared to other states. In fact, in grades 3-5, the reading cut score is in the 25th percentile range.

[1] A cut score is the minimum score a student must receive on the Arizona's Instrument to Measure Standards (AIMS) in order to be considered proficient under Arizona's accountability system.

[2] Note that Arizona received full approval from the U.S. Department of Education to implement a student growth model for the 2006-2007 school year. The current analysis, which draws on data from 2005–2006, does not in any way use or incorporate student growth model calculations.

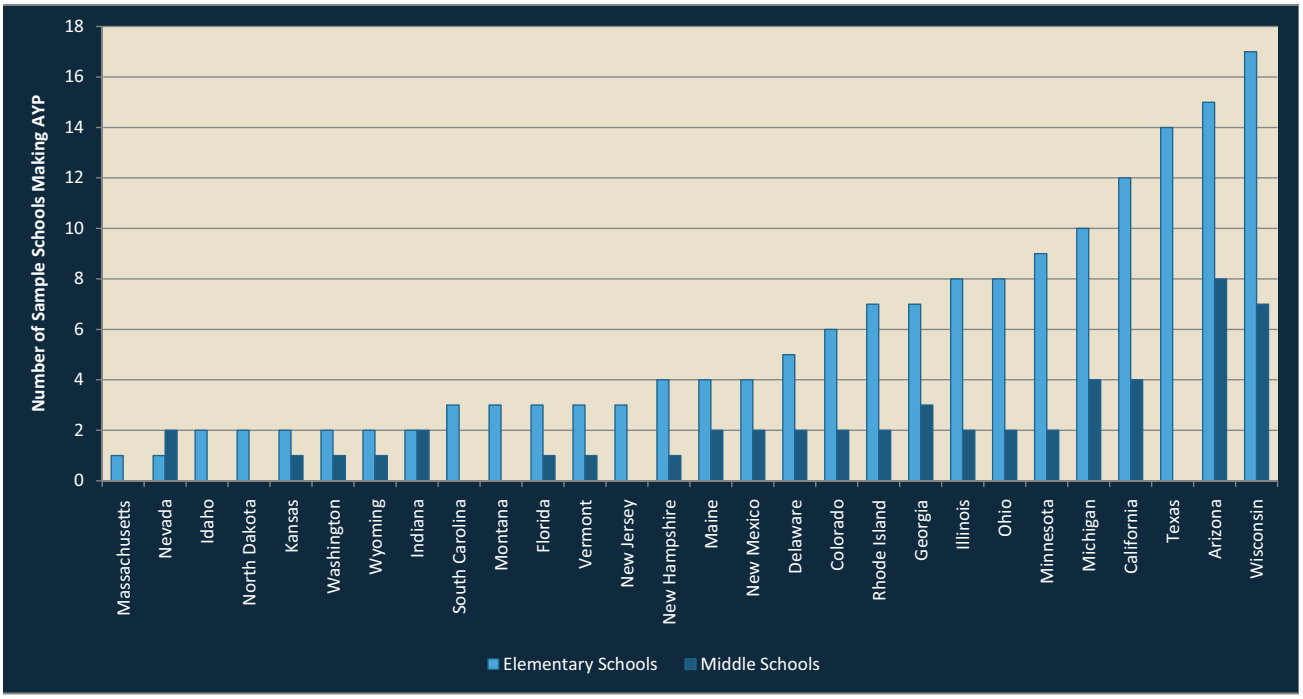

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- Nearly all of the schools in our sample that failed to make AYP in Arizona are meeting expected targets for their overall populations, but failing because of the performance of individual subgroups—particularly students with disabilities (SWDs) at the middle school level.[3]

- In Arizona, as in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As in other states, middle schools have greater difficulty reaching AYP in Arizona than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.[4]

- A strong predictor of a school making AYP under Arizona's system is whether it has enough SWDs to qualify as a separate subgroup. In cases where there were enough students to constitute a separate SWD subgroup, every school with one failed to make AYP.

Introduction

The Proficiency Illusion (Cronin, et al. 2007a) linked student performance on Arizona's test and those of 25 other states to the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This

[3] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for Limited English proficient (LEP) students and SWDs may be more negative than actual findings, 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 Arizona's Instrument to Measure Standards (AIMS), the standardized state test. 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.

[4] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools--with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others) that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error–to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners,[6] among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

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

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

[7] We gave all schools in our sample pseudonyms in this report.

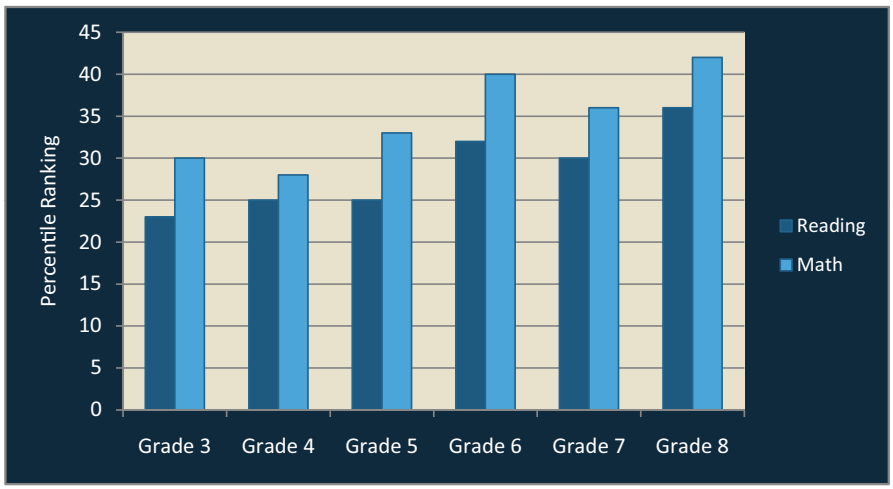

Figure 2. Arizona reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Arizona's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Arizona's cut scores are below the 45th percentile.

Proficiency cut score estimates for Arizona's Instrument to Measure Standards (AIMS) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Arizona's definitions of proficiency in reading and math were below-average to average in terms of difficulty, compared to the other states in the study. These cut scores were used to estimate whether students would have scored as proficient or better on the Arizona test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Arizona AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Arizona's 2008 AYP rules.

Table 1 shows the pertinent Arizona AYP rules that were applied to elementary and middle schools in this study. Arizona's minimum subgroup size is 40, which is comparable to most other states we examined.[8] However, the size is grade-based, meaning a school must have at least 40 individuals within a grade for that subgroup to be evaluated. Annual targets also change according to grade and subject area. The annual target for grade 3 reading, for example, is 62% of students reaching proficiency; that number changes to 38% for grade 8 math.

Furthermore, although most states apply confidence intervals (or margins of statistical error) to their measurement of student proficiency rates, Arizona's 99% confidence interval gives schools greater leniency than the 95% confidence interval used by most other states. So, for instance, although schools are supposed to get 38% of their eighth grade students to the proficient level on the state math test—and 38% of their students in each subgroup—applying the confidence interval means that the real target can actually be lower, particularly with smaller groups.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle

[8] Keep in mind that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

Table 1. Arizona AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 40	
	Low-income students: 40	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	44.0	62.6
Grade 4	45.0	56.0
Grade 5	32.0	54.6
Grade 6	45.0	56.0
Grade 7	49.0	59.2
Grade 8	31.0	54.0
MATH		
Grade 3	32.0	54.6
Grade 4	54.0	63.2
Grade 5	20.0	46.6
Grade 6	43.0	54.4
Grade 7	48.0	58.4
Grade 8	7.0	38.0

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each school's subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under Arizona's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Arizona's 2008 AYP rules. **Only 3 of the 18 elementary schools failed to make AYP under the Arizona rules.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The two schools with lowest average student performance (Clarkson and Maryweather) both fail to make AYP, as does one of the schools with higher average student

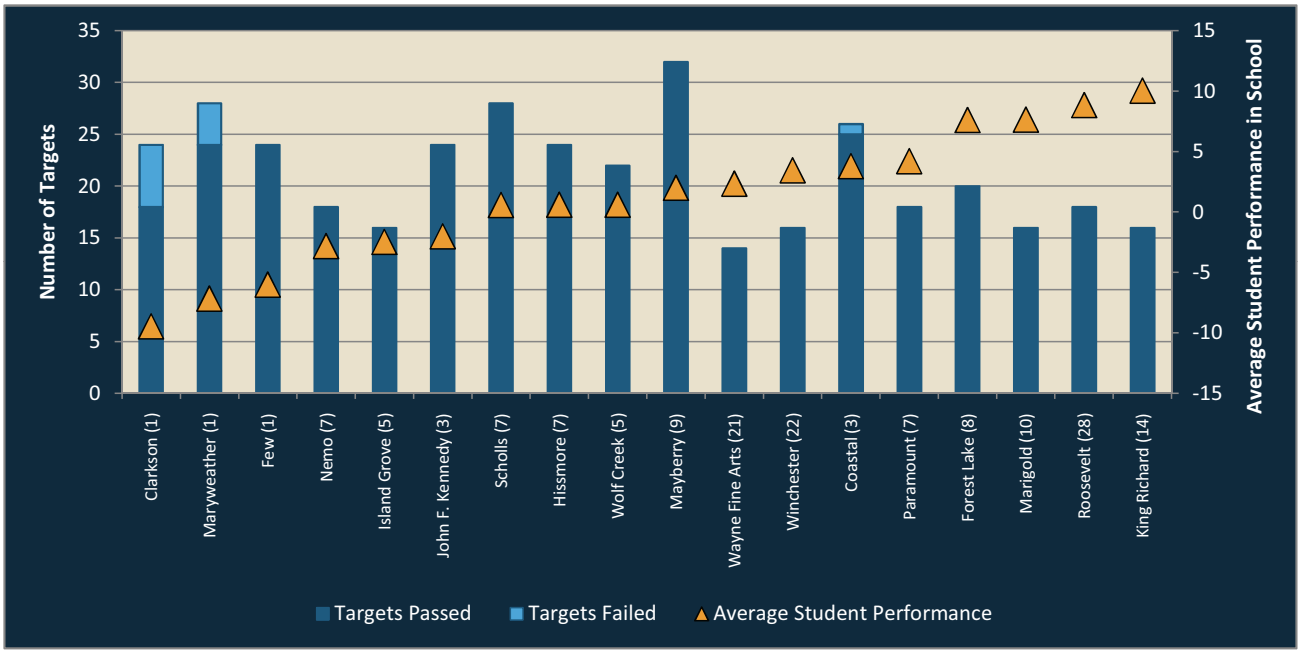

Figure 3. AYP Performance of the elementary school sample under Arizona's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Arizona's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means the school failed. Coastal Elementary, for example, met 25 of its 26 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles) which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school would have made AYP.

performance (Coastal). All three schools that failed to make it, however, have between 24 and 28 targets to meet, as opposed to the schools that made AYP, which have, on average, only 20 targets to meet.[9]

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Arizona AYP rules. **Out of 18 middle schools in our sample, 8 made AYP** – three low-performance schools (Pogesto, Chesterfield, and Filmore), and five high-performance schools (Lake Joseph, Ocean View, Walter Jones, Artemus, and Chaucer). As with the sample elementary schools, schools that made AYP tended to have fewer targets to meet than schools that didn't make AYP.

Figure 5 indicates the degree to which elementary schools' math proficiency rates are aided by the confidence interval. On this figure, the darker portions of the bars show the actual proficiency rates at each school, and the lighter portions of the bars show the degree to which these proficiency rates were "increased" by the application of the confidence interval. The orange lines show the annual measurable objective needed to meet AYP. The figure shows that none of the sample elementary schools was assisted by the confidence intervals, because the math targets in Arizona are low relative to the schools' overall performance. Although not shown, this same trend held true for middle school math and reading proficiency rates at the middle and elementary school levels as well. **Because of the relatively easy targets established by Arizona's annual measurable objectives, confidence intervals have little impact on whether schools make AYP.** [10]

[9] Recall that 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.

[10] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figure 5. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to this report's length and complexity.

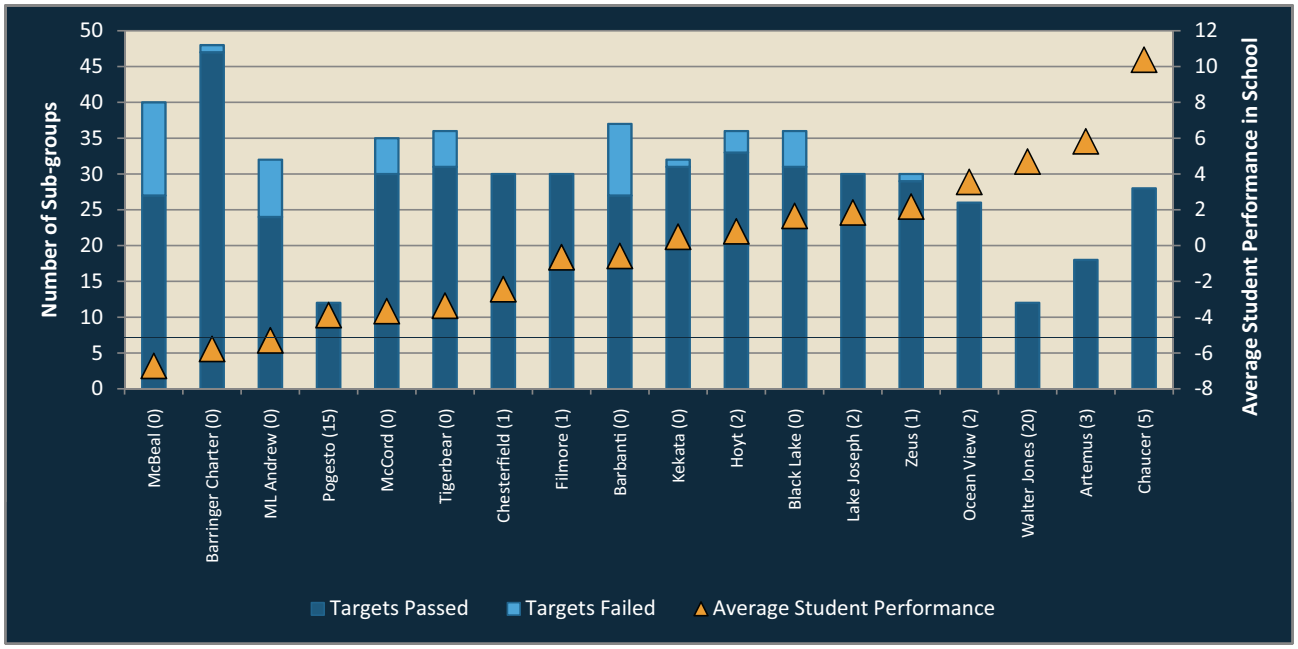

Figure 4. AYP performance of the middle school sample under Arizona's 2008 AYP rules

Note: This figure shows how each middle school would have faired under Arizona's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means the school failed. Zeus Middle School, for example, met 29 of its 30 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles) which is measured by average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school would make AYP.

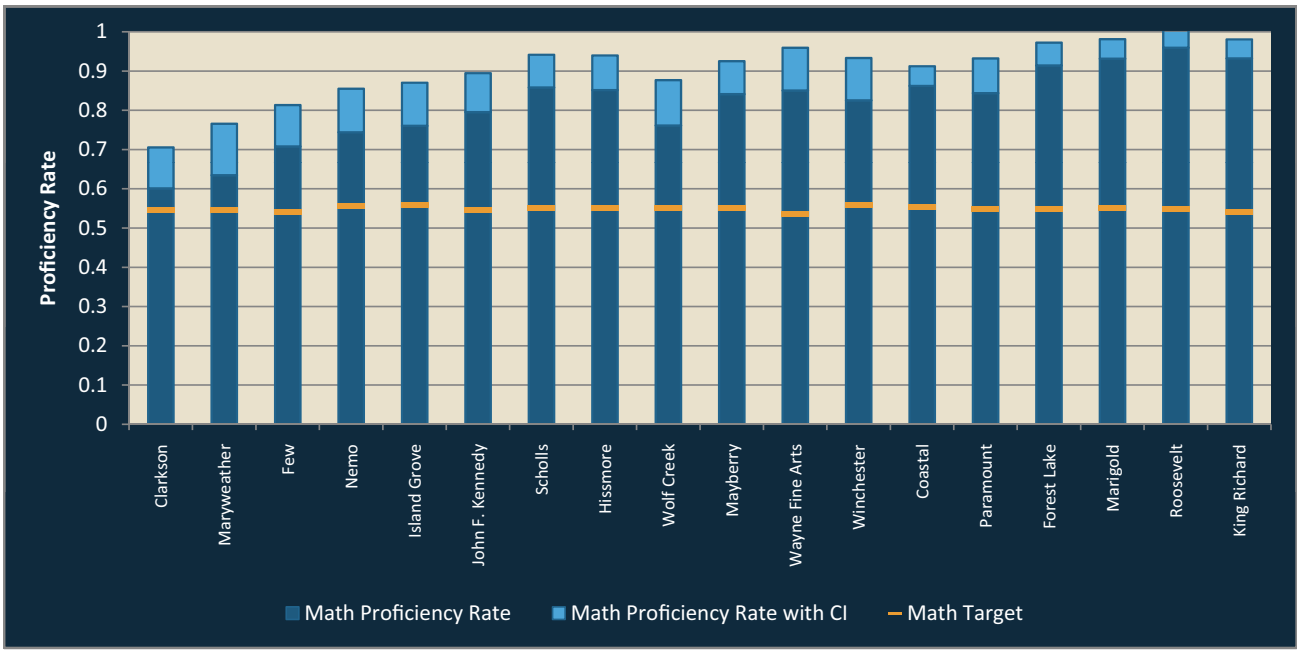

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Arizona AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	70.6%	58.1%	Y	N					Y	N					Y	N					24	18	75%	N	1
Maryweather	76.6%	68.2%	Y	Y			Y	N	Y	Y					Y	N					28	24	86%	N	1
Few	81.3%	70.6%	Y	Y					Y	Y					Y	Y					24	24	100%	Y	1
Nemo	85.5%	85.4%	Y	Y															Y	Y	18	18	100%	Y	7
Island Grove	87.0%	83.1%	Y	Y															Y	Y	16	16	100%	Y	5
JFK	89.5%	78.7%	Y	Y					Y	Y									Y	Y	24	24	100%	Y	3
Scholls	94.2%	84.7%	Y	Y			Y	Y	Y	Y									Y	Y	28	28	100%	Y	7
Hissmore	94.0%	86.7%	Y	Y					Y	Y									Y	Y	24	24	100%	Y	7
Wolf Creek	87.7%	85.3%	Y	Y			Y	Y											Y	Y	22	22	100%	Y	5
Alice Mayberry	92.5%	88.7%	Y	Y			Y	Y	Y	Y	Y	Y							Y	Y	32	32	100%	Y	9
Wayne Fine Arts	95.9%	96.4%	Y	Y															Y	Y	14	14	100%	Y	21
Winchester	93.3%	94.2%	Y	Y															Y	Y	16	16	100%	Y	22
Coastal	91.2%	85.3%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	26	25	96%	N	3
Paramount	93.2%	88.6%	Y	Y															Y	Y	18	18	100%	Y	7
Forest Lake	97.3%	94.7%	Y	Y					Y	Y									Y	Y	20	20	100%	Y	8
Marigold	98.1%	94.7%	Y	Y															Y	Y	16	16	100%	Y	10
Roosevelt	100.4%	99.7%	Y	Y															Y	Y	18	18	100%	Y	28
King Richard	98.1%	96.3%	Y	Y															Y	Y	16	16	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP. Unlike most states, Arizona schools consider each grade separately when determining whether the minimum *n* size is exceeded for a particular subgroup. This means that Arizona schools may be required to meet up to 18 targets for each grade (2 targets each–math and reading–for the overall population, SWDs, LEP, low income, African American, Asian, Hispanic, American Indian, and white). This is, of course, provided that there are sufficient numbers of students within the grade to exceed the state's minimum *n* size of 40 in every subgroup. (In actuality, it's much harder to exceed the minimum *n* size when individual grade levels are considered versus the school as a whole.) In this table, for example, we see that Clarkson Elementary met the minimum *n* size for its overall, Hispanic, and low income subgroups. However, to preserve space, each grade is not displayed separately. Consequently, the number of AYP targets required at Clarkson (24) and the number of targets met (18), let us know that the school failed to meet all of its required subgroup targets, but we don't know in which grades.

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low average student performance can still make AYP when the school has relatively few targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential sub-

Table 3. Middle school subgroup performance of sample schools under the 2008 Arizona AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	62.9%	66.0%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	40	27	68%	N	0
Barringer Charter	66.9%	69.4%	Y	Y					Y	Y	Y	N			Y	Y					48	47	98%	N	0
ML Andrew	63.9%	71.6%	Y	Y	N	N			N	Y	N	N			Y	Y			Y	Y	32	24	75%	N	0
Pogesto	77.7%	92.1%	Y	Y																	12	12	100%	Y	15
McCord Charter	65.8%	72.9%	Y	Y			Y	Y	N	N	N	Y			Y	Y			Y	Y	35	30	86%	N	0
Tigerbear	73.2%	71.5%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	36	31	86%	N	0
Chesterfield	78.4%	75.1%	Y	Y					Y	Y	Y	Y							Y	Y	30	30	100%	Y	1
Filmore	76.4%	82.2%	Y	Y					Y	Y					Y	Y			Y	Y	30	30	100%	Y	1
Barbanti	69.6%	75.0%	Y	Y	N	N		N	N	N					Y	Y			Y	Y	37	27	73%	N	0
Kekata	80.4%	77.9%	Y	Y	N	Y			Y	Y	Y	Y							Y	Y	32	31	97%	N	0
Hoyt	81.7%	80.9%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	36	33	92%	N	2
Black Lake	83.5%	80.3%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	36	31	86%	N	0
Lake Joseph	82.1%	86.5%	Y	Y					Y	Y					Y	Y			Y	Y	30	30	100%	Y	2
Zeus	83.7%	82.2%	Y	Y		N			Y	Y									Y	Y	30	29	97%	N	1
Ocean View	86.4%	91.4%	Y	Y					Y	Y					Y	Y			Y	Y	26	26	100%	Y	2
Walter Jones	100.0%	99.9%	Y	Y																	12	12	100%	Y	20
Artemus	90.3%	92.5%	Y	Y					Y	Y									Y	Y	18	18	100%	Y	3
Chaucer	91.4%	93.1%	Y	Y					Y				Y	Y	Y	Y			Y	Y	28	28	100%	Y	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP. Unlike most states, Arizona schools consider each grade separately when determining whether the minimum *n* size is exceeded for a particular subgroup. This means that Arizona schools may be required to meet up to 18 targets for each grade (2 targets each–math and reading–for the overall population, SWDs, LEP, low income, African American, Asian, Hispanic, American Indian, and white). This is, of course, provided that there are sufficient numbers of students within the grade to exceed the state's minimum *n* size of 40 in every subgroup. (In actuality, it's much harder to exceed the minimum *n* size when individual grade levels are considered versus the school as a whole.) In this table, for example, we see that Barringer Charter met the minimum *n* size for its overall, African American, Hispanic, and low income subgroups. However, to preserve space, each grade is not displayed separately. Consequently, the number of AYP targets required at Barringer Charter (48) and the number of targets met (47), let us know that the school failed to meet all of its required subgroup targets, but we don't know in which grades.

groups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Arizona rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- No elementary schools failed to meet their overall targets for math.
- One elementary school (Clarkson) failed to meet the overall target for reading.
- All middle schools met overall targets for reading and math.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Arizona AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	1	0	1
Students with limited English proficiency	4	0	1
Low-income students	9	0	1
African-American students	2	0	0
Asian/Pacific Islander students	0	0	0
Hispanic students	3	0	2
American Indian/Alaska Native students	0	0	0
White students	15	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Arizona AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	7	7
Students with limited English proficiency	3	1	2
Low-income students	16	4	3
African-American students	8	2	2
Asian/Pacific Islander students	1	0	0
Hispanic students	9	1	1
American Indian/Alaska Native students	0	0	0
White students	15	0	0

- One elementary school (Coastal) met every target except for the reading target for its SWDs.
- Five middle schools (Tigerbear, Kekata, Hoyt, Black Lake, and Zeus) met all targets except for SWDs.
- One middle school (Barringer Charter) met every target except for one ethnic minority group.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. As shown, the performance of SWDs is proving most challenging for schools under Arizona's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum n of 40. In fact, every school within the sample with qualifying SWDs failed to make AYP. (However, it's well worth noting that only one school met the minimum *n* size for SWD subgroups at the elementary level.)

Table 6. Comparisons between schools that did and didn't make AYP in Arizona, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	15	3	8	10
Average student body size	299	333	587	1077
Average % low income	41	75	34	54
Average % nonwhite	34	72	43	45
Average performance†	2.32	-4.26	2.41	-2.03
Average % growth‡	118	100	106	92
Average number of targets to meet	20	26	23	36

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Arizona's NCLB accountability system is, in some respects, behaving similarly to those in other states. All the sample schools that fail under Arizona rules failed in most of the other states examined in this study. For example, among the elementary schools in our sample, Clarkson and Maryweather both failed in Arizona (Figure 3), and these two schools failed in all but one of the 28 states examined in this study. Likewise, all the failing middle schools in Figure 4 also failed in the majority of the other states examined in the study.

However, on the whole, Arizona's AYP rules are generally more lenient than in other states. Many sample elementary schools (e.g., Few, Island Grove, and JFK) and middle schools (e.g., Chesterfield and Filmore) that failed to make AYP in most other states make it in Arizona. This is most likely attributable to Arizona's minimum subgroup policy, which considers grades separately, meaning that an Arizona school will have fewer accountable subgroups than a similar school in another state. Arizona's subgroup policies, along with relatively easy annual targets relative to student performance, mean that schools made AYP more easily in Arizona than in many other states.

Despite its greater leniency, the rule set in Arizona showed certain trends that were similar for other states as well. Schools that made AYP in Arizona tended to have higher average student performance than schools that didn't, though schools with more targets to meet tended not to do as well as schools with fewer targets.

This is illustrated in Table 6, which compares schools that did and didn't make AYP on a number of academic and demographic dimensions in Arizona. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have smaller student populations, particularly in middle schools, fewer subgroups (and thus fewer targets to meet), and lower percentages of low income students.

Concluding Observations

This study evaluated the test performance data of students from 18 elementary and 18 middle schools across

the country to see how these schools would fare under Arizona's AYP rules (and AMOs) for 2008. We found that 15 elementary schools and 8 middle schools—23 in all, from a sample of 36—would have made AYP in Arizona. Compared to the other 27 states examined, this places Arizona at the high end of the distribution in terms of the number of schools making AYP (see Figure 1). In addition, some sample schools make AYP in Arizona that fail to make AYP in most other states. This is most likely because Arizona's proficiency standards are relatively easy compared to other states and its particular rules result in fewer accountable subgroups.

Because the overriding goal of the federal NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Arizona's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students such as low-income or Hispanic students. All the sample schools, save one, make AYP in Arizona for their student populations as a whole (i.e., without considering sub-group results). In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on. has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Is it "fair" for a state to have such generous margins of error and low elementary school cut scores? Does it make sense that the size of a school's enrollment has so much influence over making AYP? These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

California

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines California's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough California's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under California's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against California's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **6 of 18 elementary schools** and **14 of 18 middle schools** in our sample failed to make AYP in 2008 under California's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, **we find that only three states exceeded California in terms of the number of elementary schools making AYP (Texas, Arizona, and Wisconsin).**
- In California, subgroups of students (such as minorities or low-income children[2]) must be quite large in order to be counted separately in AYP calculations. In this way, **the achievement scores of many minority, disabled, or limited English proficient students that do not count separately in California would count separately in most of the other states.**
- Furthermore, although the majority of states examined in the study apply confidence intervals (or margins of statistical error) to their student proficiency rates, California's 99% confidence interval gives schools greater leniency than the 95% confidence interval used by most other states. Such a lenient confidence interval might normally rescue otherwise failing schools, but because California's minimum subgroup size is rather large anyway and because the state places limitations on the use of intervals, it is seldom used.

More schools in the study make AYP in **California** than in most other states. There are several factors which contribute to this. First, though California has relatively high proficiency standards (or cut scores) in reading and math, the percentage of students required to meet those standards in 2008 is relatively low (roughly 35 percent proficient in English Language Arts and 37 percent proficient in math). An additional factor is that the minimum subgroup size for reporting purposes is relatively high. California's minimum subgroup size is generally 100 students (they also use a "sliding" *n* size depending on a school's enrollment). This is larger than the minimum subgroup size used by most other states examined in the study. Hence, the achievement scores of many minority, limited English proficient (LEP), and disabled students that are *not* counted separately for accountability purposes in California would be counted separately in most other states.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the California Standards Test.

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

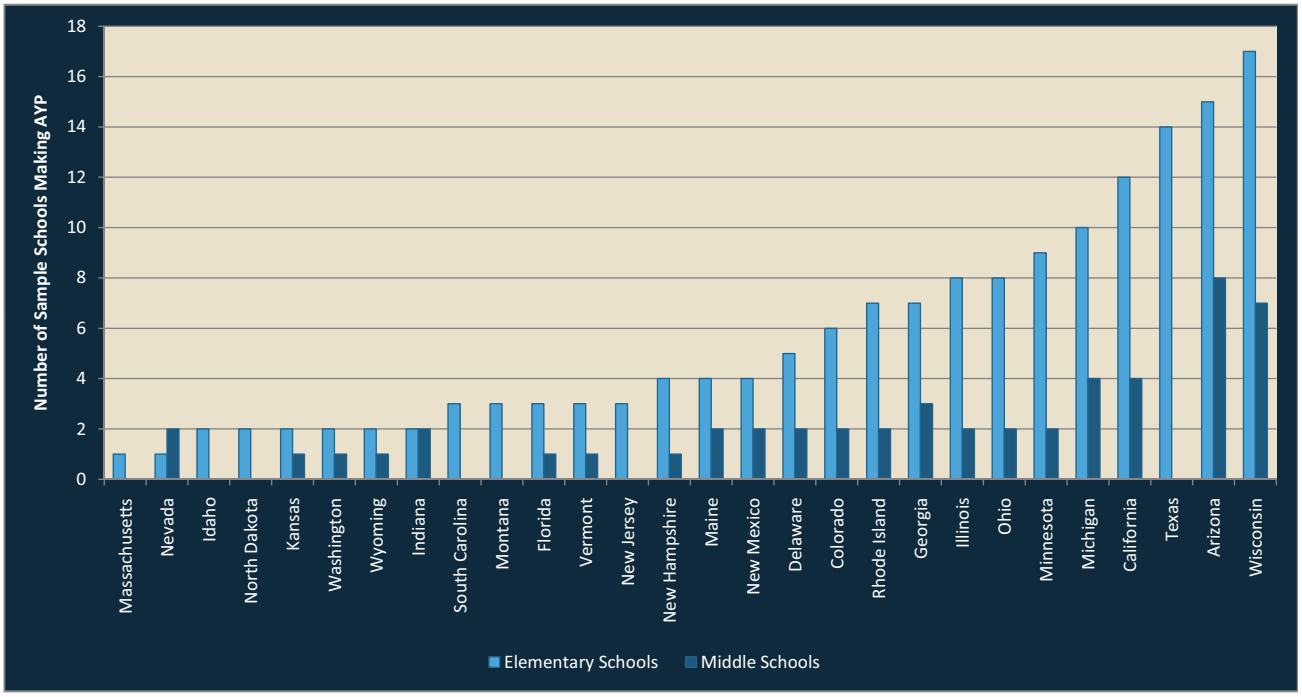

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- **California's accountability system, then, has high cut scores and high minimum subgroup sizes, but fairly low annual targets (hovering around proficiency levels of 35%).**
- Still, many of the schools in our sample that failed to make AYP in California did meet expected targets for their overall populations but failed because of the performance of individual subgroups.[3]
- In California, as in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.
- As in other states, middle schools have greater difficulty reaching AYP in California than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.
- A strong predictor of a school making AYP under California's system is whether it has enough English language learners to qualify as a separate subgroup. Almost every single school with a subgroup of students with limited English proficiency (LEP)[4] failed to make AYP. Likewise, most of the schools with enough qualifying students with disabilities (SWDs) failed to meet their AYP targets.[5]

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

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

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 California Standards Test, the standardized state test. 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.

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on California's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

[6] We gave all schools in our sample pseudonyms in this report.

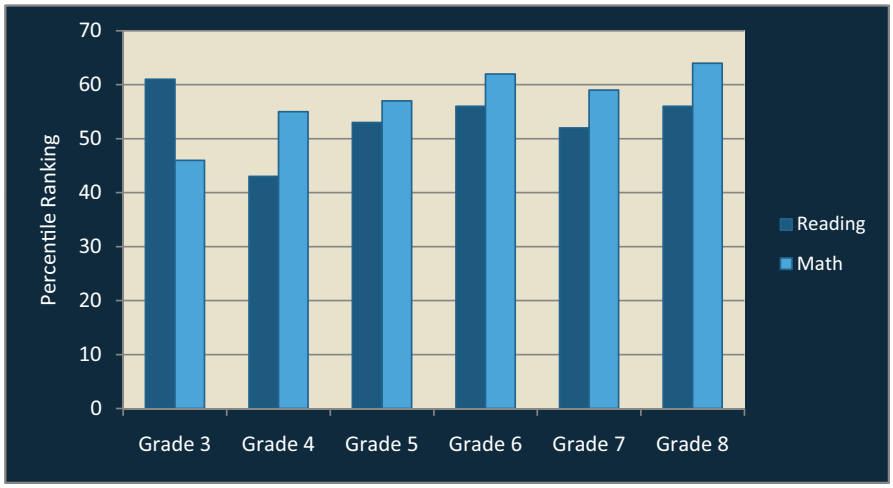

Figure 2. California reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of California's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of California's cut scores are at or below the 65th percentile.

Proficiency cut score estimates for the California Standards Tests (CST) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that California's definitions of proficiency in reading and math were relatively difficult compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the California test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under California AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against California's 2008 AYP rules.

Table 1 shows the pertinent California AYP rules that we applied to elementary and middle schools in this study. California's minimum subgroup size is 15% of the student population; however, the minimum subgroup size can't be less than 50 or more than 100.[7] This is larger than the minimum subgroup size used by most other states examined in the study.

Furthermore, although the majority of states examined in the study apply confidence intervals (or margins of statistical error) to their student proficiency rates, California's 99% confidence interval gives schools greater leniency than the 95% confidence interval used by most other states. So, for instance, although schools are supposed to get 35.2% of their grade 3-8 students (and 35.2% of their grade 3-8 students in each subgroup) to the proficient level on the state reading test, applying the confidence interval means that the real target can actually be lower. Such a lenient confidence interval might normally rescue otherwise failing schools, but two factors prevent the interval from being used that often: 1) California's minimum *n* size is rather large anyway, so fewer subgroups are held separately accountable in the first place; and 2) it is only used if the school population is fewer than 100 students.[8] **California's accountability system, then, has high cut scores and high minimum n sizes, but lenient confidence intervals and fairly low annual targets (hovering around proficiency levels of 35%).**

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision per-

[7] In California, the minimum subgroup size is 15% of the total school population. Generally, this means that the subgroup size grows with the school size. However, there's also a clause that specifies that the minimum subgroup size can't be less than 50 or more than 100. For example, a school with a total population of 500 would have a minimum subgroup size of 75 (i.e, 15%), but a school with only 300 students would have a minimum subgroup size of 50 since 15% of 300 (i.e., 45) is below the required minimum. Similarly, a school with 800 students would have a minimum subgroup size of 100, since 15% of 800 (i.e., 120) is greater than the maximum size of 100.

[8] We conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

Table 1. California AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 15% of the student population but with a minimum of 50 and maximum of 100	
	SWDs: 15% of the student population but with a minimum of 50 and maximum of 100	
	Low-income students: 15% of the student population but with a minimum of 50 and maximum of 100	
	LEP students: 15% of the student population but with a minimum of 50 and maximum of 100	
CI	**Applied to proficiency rate calculations?**	**Additional notes**
	Yes; 99% CI	Used only if school population is fewer than 100 students; not used otherwise
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	13.6	35.2
Grade 4	13.6	35.2
Grade 5	13.6	35.2
Grade 6	13.6	35.2
Grade 7	13.6	35.2
Grade 8	13.6	35.2
MATH		
Grade 3	16.0	37.0
Grade 4	16.0	37.0
Grade 5	16.0	37.0
Grade 6	16.0	37.0
Grade 7	16.0	37.0
Grade 8	16.0	37.0

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

mits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

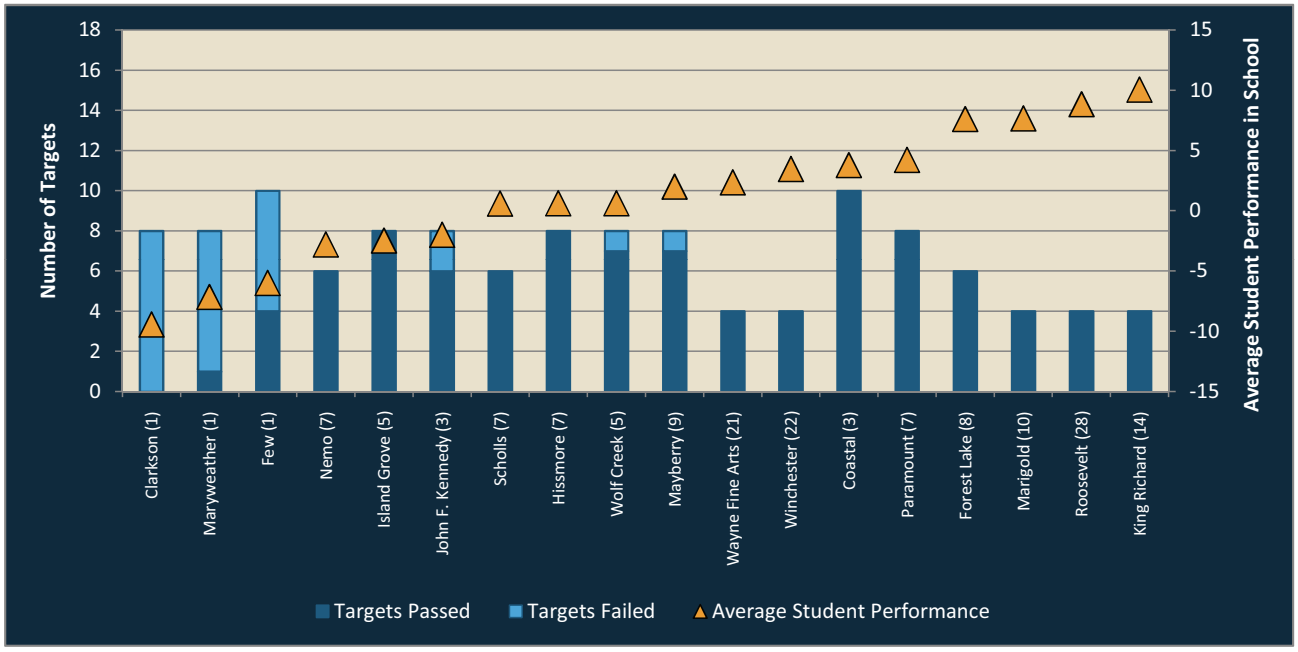

Figure 3. AYP performance of the elementary school sample under California's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under California's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Mayberry Elementary, for example, met 7 of its 8 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

How Did the Sample Schools Fare under California's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under California's 2008 AYP rules. **Twelve elementary schools made AYP and six failed to make it.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The majority of the schools making AYP are in the right half of the figure, meaning that the highest performing students were found at these schools.

Yet almost without regard to average student performance, the schools that made AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (since each subgroup has its own separate targets). For example, Wayne Fine Arts and Winchester passed, but had only four targets each. Each school must make AYP for its overall student population in reading and math (two targets) and for its white population, resulting in four total targets.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 California AYP rules. **Of 18 middle schools in our sample, only 4 made AYP**—one low-performance school (Pogesto), and three high-performance schools (Walter Jones, Artemus, and Chaucer), most of which have relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by California's confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the annual measurable objective

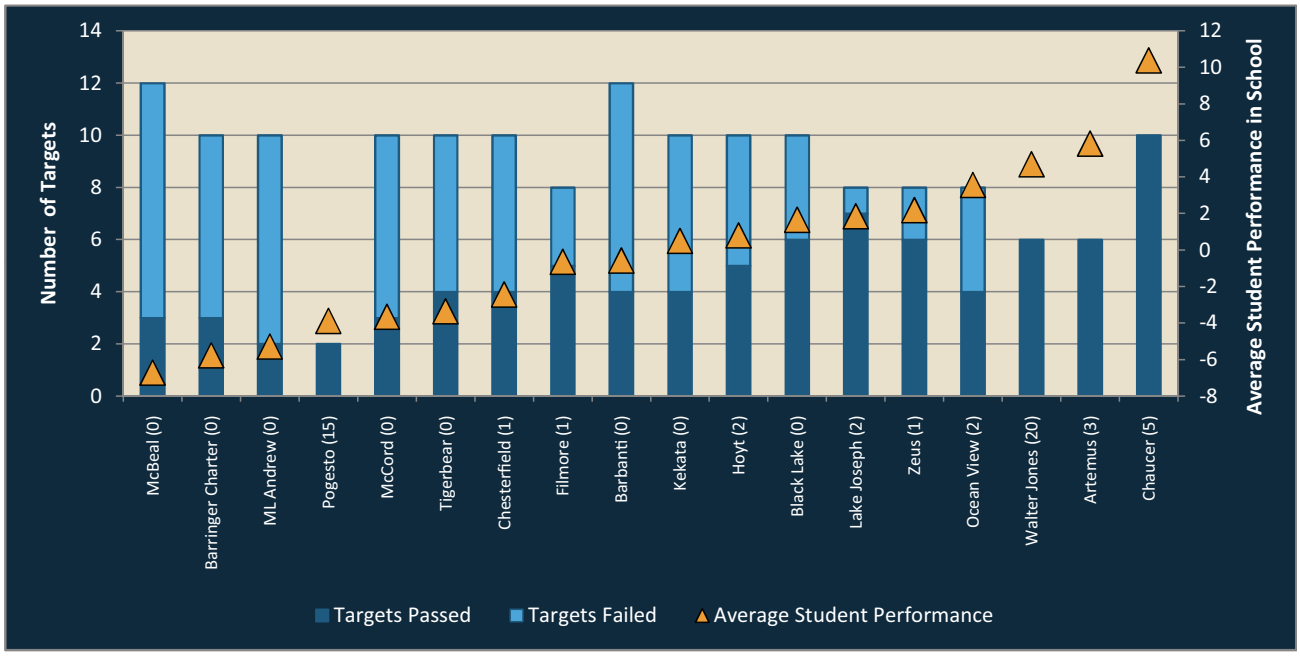

Figure 4. AYP performance of the middle school sample under California's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under California's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Lake Joseph Middle School, for example, met 7 of its 8 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

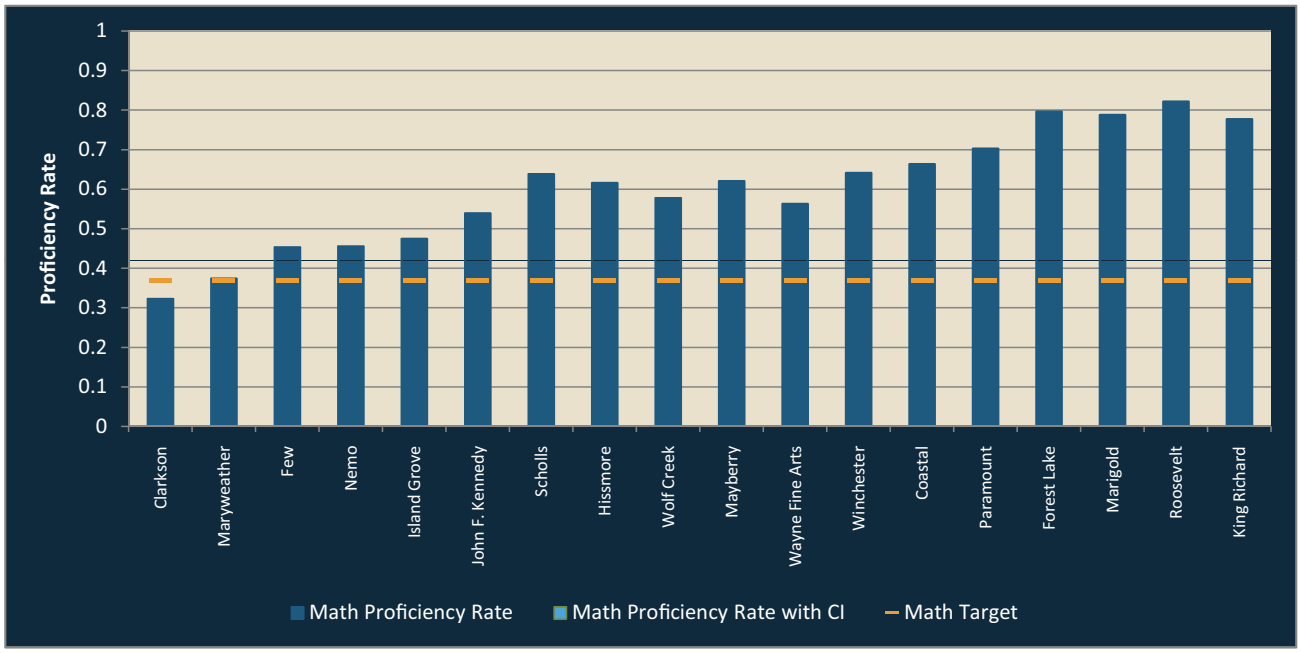

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved. Since California only makes use of the confidence interval in schools with fewer than 100 students, confidence intervals are not shown in Figure 5 (all schools have more than 100 students). If confidence intervals were used, however, they would be depicted in a lighter shade of blue on top of the dark blue bar. Annual targets are indicated by the orange lines.

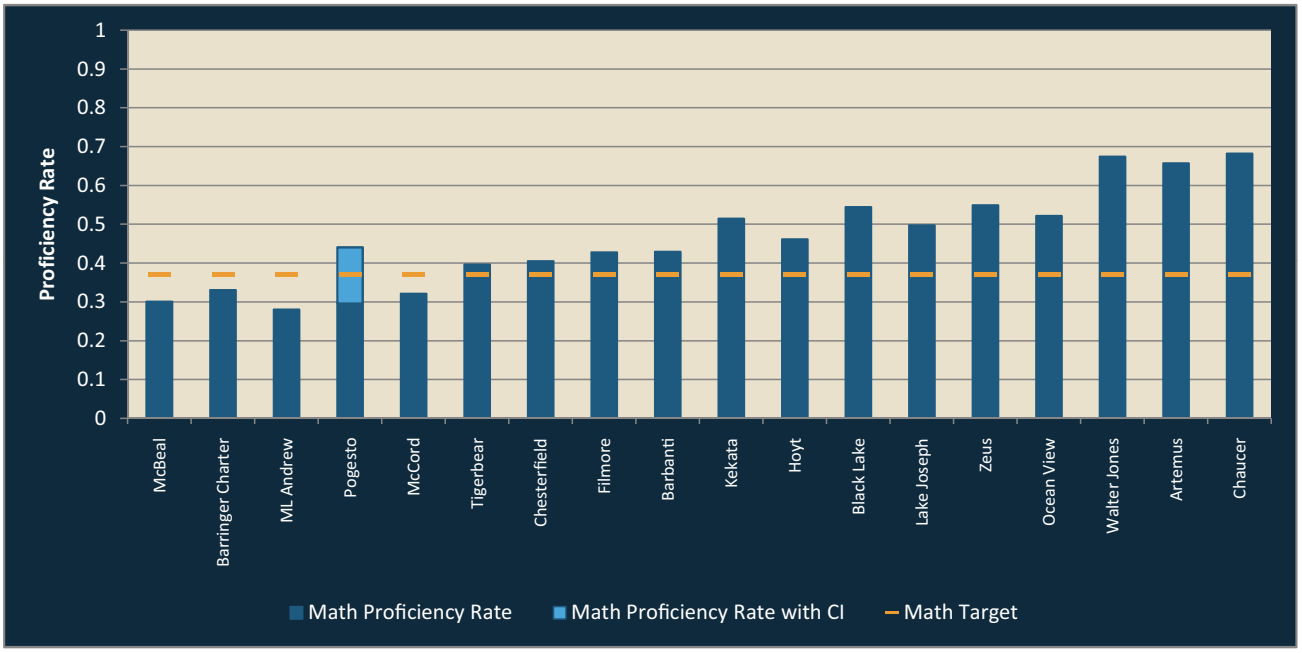

Figure 6. Impact of the confidence interval on middle school math proficiency rates

NNote: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. Since California only makes use of the confidence interval in schools with fewer than 100 students, confidence intervals are not shown for the most part. Pogesto, in fact, is the only eligible school and the only one that meets its overall math target via the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

needed to meet AYP. These figures show that no sample elementary schools and one middle school (Pogesto) were assisted by the confidence interval. It's important to keep in mind, however, that Pogesto was the only school *eligible* to make use of the confidence interval (California rules allow confidence intervals to be used only in schools with fewer than 100 students.)

The effect of confidence intervals on reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, no elementary schools and only one middle school (Pogesto again) met the overall targets with the confidence interval. In short, **the application of the confidence interval had little or no impact on whether the sample elementary and middle schools met California's overall reading and math targets.**[9] **So, even though we would expect California's generous confidence interval to rescue otherwise failing schools, we see that the state's high minimum n size and low school enrollment requirement prevent the interval from serving that function.**

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American In-

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary subgroup performance of sample schools under the 2008 California AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	32.3%	18.6%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	37.4%	32.9%	Y	N			N	N	N	N					N	N					8	1	13%	N	1
Few	45.4%	32.6%	Y	N	N	N	Y	N	Y	N					Y	N					10	4	40%	N	1
Nemo	45.6%	44.7%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	47.5%	50.6%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	4
JFK	53.9%	42.2%	Y	Y					Y	N	Y	N							Y	Y	8	6	75%	N	3
Scholls	63.8%	48.0%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Hissmore	61.6%	50.0%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Wolf Creek	57.8%	54.3%	Y	Y					Y	Y					Y	N			Y	Y	8	7	88%	N	5
Alice Mayberry	62.1%	50.9%	Y	Y					Y	Y	Y	N							Y	Y	8	7	88%	N	9
Wayne Fine Arts	56.3%	61.5%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	64.2%	63.0%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	66.3%	63.0%	Y	Y	Y	Y			Y	Y	Y	Y							Y	Y	10	10	100%	Y	3
Paramount	70.3%	62.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	7
Forest Lake	79.6%	70.4%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	8
Marigold	78.8%	76.5%	Y	Y															Y	Y	4	4	100%	Y	10
Roosevelt	82.2%	79.0%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	77.7%	82.3%	Y	Y															Y	Y	4	4	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

dian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 California rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that

- One elementary school (Clarkson) and four middle schools (McBeal, Barringer Charter, ML Andrew, and McCord Charter) failed to meet math targets for their overall school populations.
- One elementary school (Few) and nine middle schools failed the AMOs for their SWDs.
- All elementary schools (Clarkson, Maryweather, and Few) and middle schools (McBeal, Barbanti) with qualified LEP subgroups failed to make AYP.
- Four elementary schools and nine middle schools failed to meet the AMOs for low-income students.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. As shown, **California's minimum *n* of 100 means that the schools in the sample have essentially five subgroups—SWDs, low-income, Hispanic/Latino, African American, and**

Table 3. Middle school subgroup performance of sample schools under the 2008 California AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	30.1%	36.2%	N	Y	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Barringer Charter	33.1%	35.3%	N	Y	N	N			N	N	N	N			Y	Y					10	3	30%	N	0
ML Andrew	28.0%	38.1%	N	Y					N	N	N	N			N	N			N	Y	10	2	20%	N	0
Pogesto	29.6%	33.3%	Y	Y																	2	2	100%	Y	15
McCord Charter	32.2%	43.6%	N	Y					N	N	N	N			N	N			Y	Y	10	3	30%	N	0
Tigerbear	39.6%	39.0%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	0
Chesterfield	40.5%	38.2%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	1
Filmore	42.8%	47.2%	Y	Y					N	Y					N	N			Y	Y	8	5	63%	N	1
Barbanti	42.9%	45.3%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	0
Kekata	51.5%	47.3%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	0
Hoyt	46.2%	49.8%	Y	Y	N	N			N	Y	N	N							Y	Y	10	5	50%	N	2
Black Lake	54.4%	48.7%	Y	Y	N	N			Y	Y	N	N							Y	Y	10	6	60%	N	0
Lake Joseph	49.8%	53.7%	Y	Y					Y	Y					N	Y			Y	Y	8	7	88%	N	2
Zeus	54.9%	53.2%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	1
Ocean View	52.2%	63.6%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	2
Walter Jones	67.4%	66.9%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	65.7%	66.2%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	3
Chaucer	68.2%	73.9%	Y	Y					Y	Y			Y	Y	Y	Y			Y	Y	10	10	100%	Y	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

White—with sufficient numbers of students for reporting purposes. Of these subgroups, the performance of low-income students (and to a lesser extent, SWDs) is proving most challenging for schools under California's system. This is especially true in middle schools, which are generally larger and more likely to have enough students to meet the minimum *n* in the subgroups.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that California's NCLB accountability system is, in many respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in California, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in California.

But California is also home to a few anomalies. First, consider Coastal Elementary (see Figure 3). It failed to

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 California AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	2	1	1
Students with limited English proficiency	3	2	3
Low-income students	13	2	4
African-American students	4	0	2
Asian/Pacific Islander students	0	0	0
Hispanic students	6	2	4
American Indian/Alaska Native students	0	0	0
White students	15	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 California AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	9	9	9
Students with limited English proficiency	2	2	2
Low-income students	17	11	9
African-American students	8	8	8
Asian/Pacific Islander students	1	0	0
Hispanic students	9	7	6
American Indian/Alaska Native students	0	0	0
White students	16	1	0

make AYP in 25 of the 28 states in our sample, yet made AYP in California. In examining Table 2, we can see that Coastal didn't meet the minimum numbers for the LEP or Hispanic subgroups, which created difficulty for many schools in the study. Without those particular subgroups counting, Coastal was able to meet AYP, even when it failed under the standards of most other states.

This is consistent with the patterns shown in Table 6, which compares the schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower percentages of low-in-

Table 6. Comparisons between schools that did and didn't make AYP in California, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	12	6	4	14
Average student body size	262	390	520	956
Average % low income	35	70	28	50
Average % nonwhite	30	62	27	49
Average performance†	3.67	-3.66	4.25	–1.29
Average % growth‡	121	103	121	92
Average number of targets to meet	6	8	6	9

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

come and nonwhite students. Similarly, middle schools that made AYP have slightly higher performing students, on average, than middle schools that didn't make it, but have smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

This study examined the test performance data of students in 18 elementary and 18 middle schools across the country to see how those schools would fare under California's AYP rules (and AMOs) for 2008. We found that 12 elementary schools and 4 middle schools—16 in all, from a sample of 36—would have made AYP in California. Looking across the 28 state accountability systems examined in the study, this places California at the high end of the distribution in terms of the number of schools making AYP (see Figure 1).

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, California's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Almost all of the sample schools made AYP in California for their student populations as a whole (i.e., without considering subgroup results). In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? In the case of California, does it make sense that high cut scores can be

"tamed" by low annual targets,[10] or that large minimum *n* sizes mean that the achievement scores of students with disabilities or limited English proficiency are not counted separately? These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[10] There is some evidence that California is now rapidly increasing its annual targets. So even though the current accountability system has its drawbacks, California appears to be trying to remedy and align its various components..

Colorado

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Colorado's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Colorado's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Colorado's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Colorado's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **12 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under Colorado's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students.)
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in Colorado was exceeded in 10 other sample states. In addition, Colorado was one of 10 states with two passing middle schools in the sample (see Figure 1).**
- Most of the schools in our sample that failed to make AYP in Colorado are meeting expected targets for their overall populations but failing because of the performance of individual subgroups, particularly students with disabilities (SWD)[2] and English language learners.[3]

Colorado is a state with an interesting set of rules, which, when working in tandem, put the state in the middle of the sample distribution in terms of how many schools make AYP. First, Colorado's proficiency standards (or cut scores) are relatively easy to achieve. All of them are at or below the 25th percentile in both reading and math. Still, while Colorado's cut scores are low, its annual targets for proficiency–which vary depending on subject and grade–are fairly ambitious (ranging from 79 to 88 percent in 2008); thus, some schools do not make AYP in Colorado *despite* its undemanding proficiency standards. Another wrinkle is that Colorado's minimum subgroup size is 30, smaller than most other states we examined. This means that schools in Colorado will have more subgroups to account for than schools in most other states. In Colorado, then, schools large enough to have many accountable subgroups fail to make AYP while very small, homogenous schools tend to make AYP, even if their overall student achievement is lower.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Colorado Student Assessment Program (CSAP).

[2] SWDs are defined as those students following individualized education plans.

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

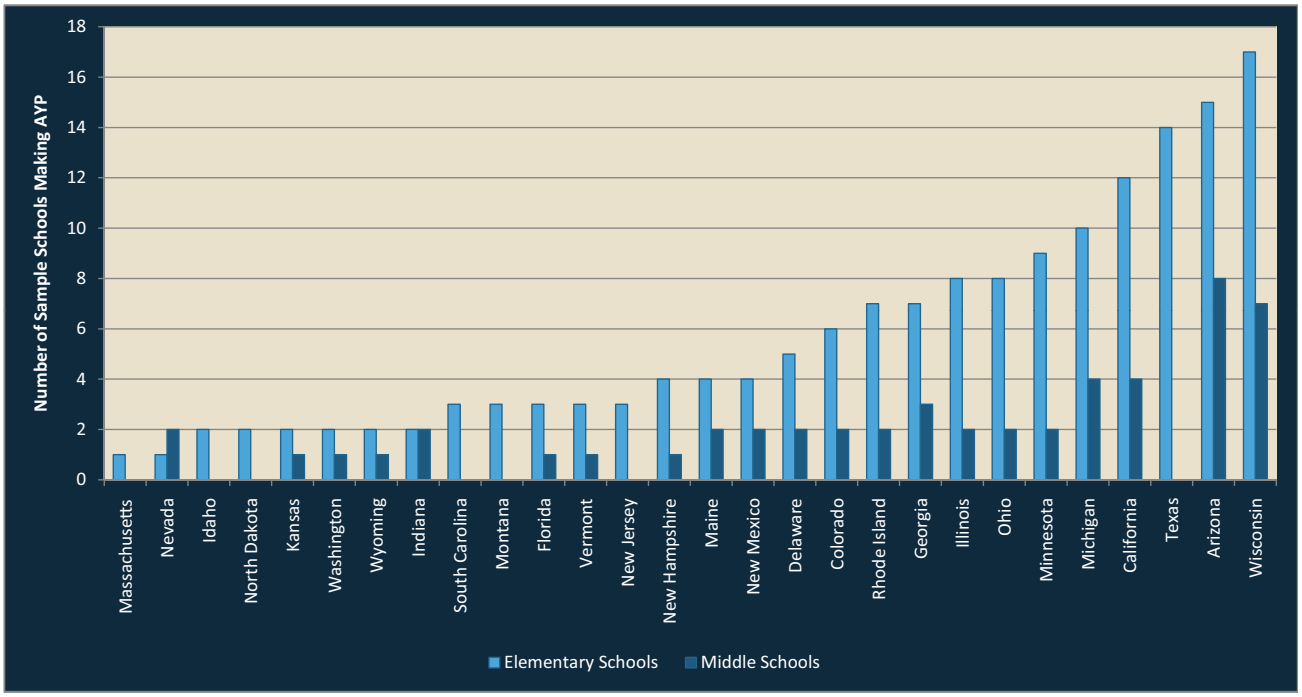

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- One sample school that failed to make AYP in most other states made AYP in Colorado. This is probably because Colorado's proficiency standards (or cut-off scores) are relatively easy compared to other states; this school also had fewer accountable subgroups.

- **Still, while Colorado's proficiency standards are low, its annual targets for proficiency are fairly ambitious (ranging from 79 to 88 percent in 2008); thus, large numbers of schools do not make AYP in Colorado *despite* its undemanding proficiency standards.**

- In Colorado, as in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- In Colorado, as in most states, middle schools have greater difficulty reaching AYP than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower.

- A strong predictor of a school making AYP under Colorado's system is whether it has enough limited English proficient (LEP) students[4] to qualify as a separate subgroup. Almost every single school with even one such subgroup failed to make AYP.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Colorado's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a

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

[5] We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Colorado Student Assessment Program, the standardized state test. 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.

computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013-2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Colorado, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the Colorado Student Assessment Program (CSAP) are taken from *The Profi-*

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

[7] We gave all schools in our sample pseudonyms in this report.

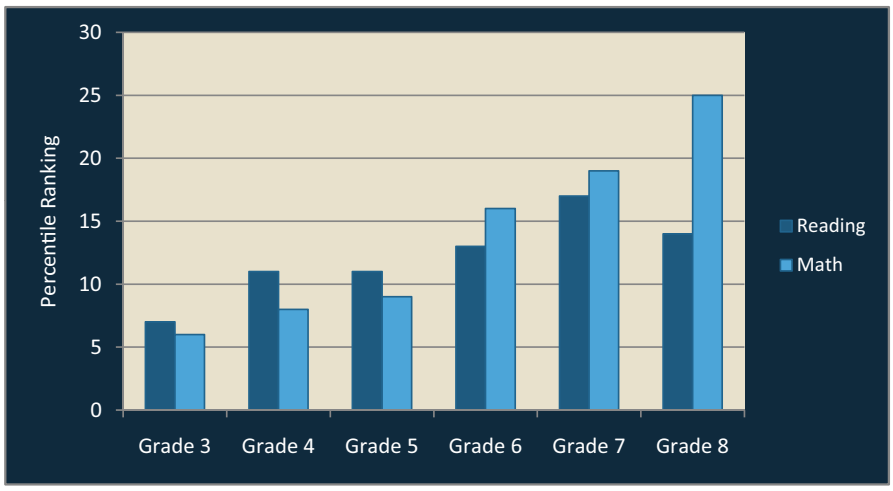

Figure 2. Colorado reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Colorado's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Colorado's cut scores are at or below the 25th percentile.

ciency Illusion (as shown in Figure 2), which found that Colorado's definitions of proficiency ranked well below the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Colorado test, given their performance on MAP.[8] Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Colorado AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Colorado's 2008 AYP rules.

Table 1 shows the pertinent Colorado AYP rules that were applied to elementary and middle schools in this study. Colorado's minimum subgroup size is 30, smaller than most other states we examined.[9] This means that schools in Colorado will have more subgroups to account for than schools in most other states.

Furthermore, most states also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. Colorado, like most other states in the study, uses a 95% confidence interval. This means even though the AMO might require a school to attain, for instance, 88.4% reading proficiency among its grade 3 students, and 88.4% reading proficiency among its grade 3 students in each subgroup, the real target can be lower, particularly with smaller groups. Note, too, that for different grades and subjects, Colorado applies different AMOs, although all are relatively demanding for 2008.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states in-

[8] NCLB requires three levels of proficiency: basic, proficient, and advanced. Colorado uses four levels of proficiency on its state test (unsatisfactory, partially proficient, proficient, and advanced). In order to comply with NCLB guidelines, Colorado merged the "partially proficient" and "proficient" categories for AYP purposes. Thus, "partially proficient" students in Colorado are considered "proficient" in terms of AYP accounting. Colorado, however, continues to report four categories of proficiency in its state reporting of CSAP results.

[9] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

Table 1. Colorado AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 30	
	Low-income students: 30	
	LEP students: 30	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	77.5	88.4
Grade 4	77.5	88.4
Grade 5	77.5	88.4
Grade 6	74.6	86.8
Grade 7	74.6	86.8
Grade 8	74.6	86.8
MATH		
Grade 3	79.5	89.0
Grade 4	79.5	89.0
Grade 5	79.5	89.0
Grade 6	60.7	79.7
Grade 7	60.7	79.7
Grade 8	60.7	79.7

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

clude attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each school's subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Colorado's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Colorado's 2008 AYP rules. **Six elementary schools made AYP while 12 failed to make it.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. Most schools making

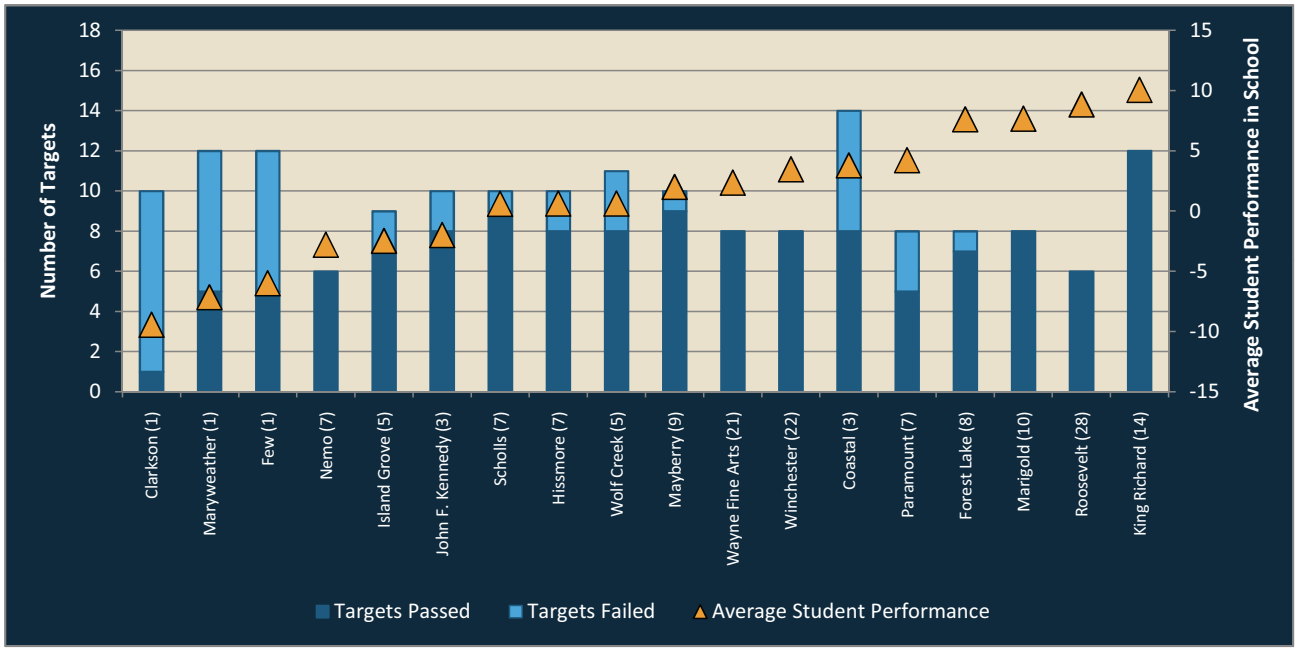

Figure 3. AYP performance of the elementary school sample under Colorado's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Colorado's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means the school fails. Mayberry Elementary, for example, met nine of its ten targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

AYP are in the right half of the figure, meaning that the higher performing students were found at these schools.

Yet almost without exception, the only schools actually to make AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (since each subgroup has its own separate targets to meet). For example, Nemo and Roosevelt made AYP, but have only six targets each.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Colorado AYP rules. **Out of 18 middle schools in our sample, only 2 made AYP** – one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by Colorado's confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the annual measurable objective needed to meet the targets. These figures show that only two elementary schools (Clarkson and Maryweather) and one middle school (Pogesto) were assisted by the confidence intervals. However, we know from Figure 3 that Clarkson and Maryweather still failed to make AYP because of low subgroup performance.

The effect of confidence intervals on reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, no elementary school is assisted by the confidence interval, but one middle school (Kekata) is helped. However, like Maryweather, Kekata failed to make AYP because of poor subgroup perform-

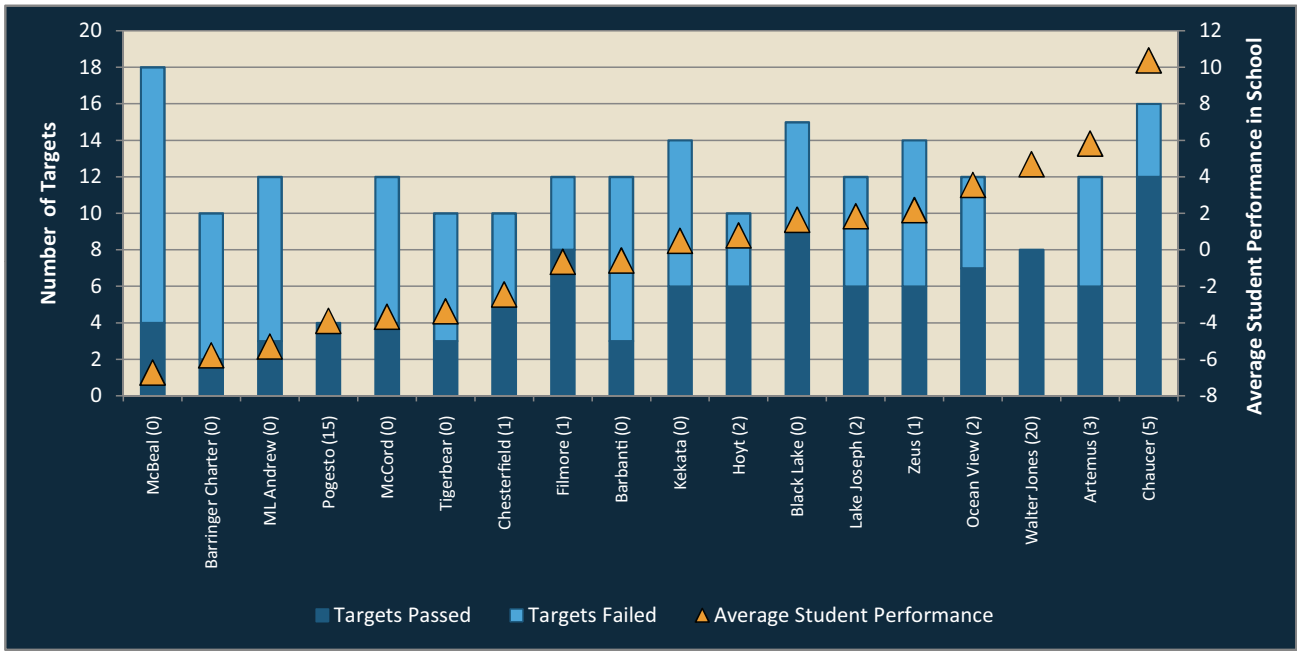

Figure 4. AYP performance of the middle school sample under Colorado's 2008 AYP rules

Note: This figure shows how each middle school within the sample would have faired under Colorado's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means that the school failed. Hoyt, for example, met 6 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

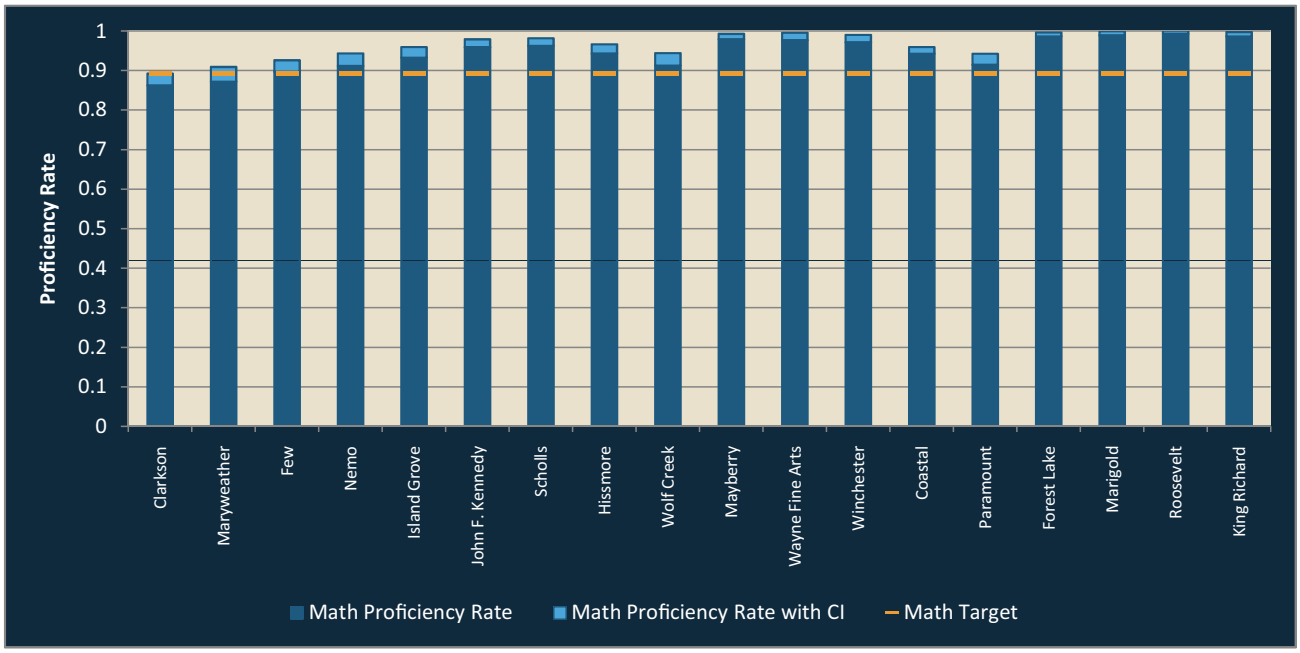

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under the Colorado 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample elementary schools, Clarkson and Maryweather, were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

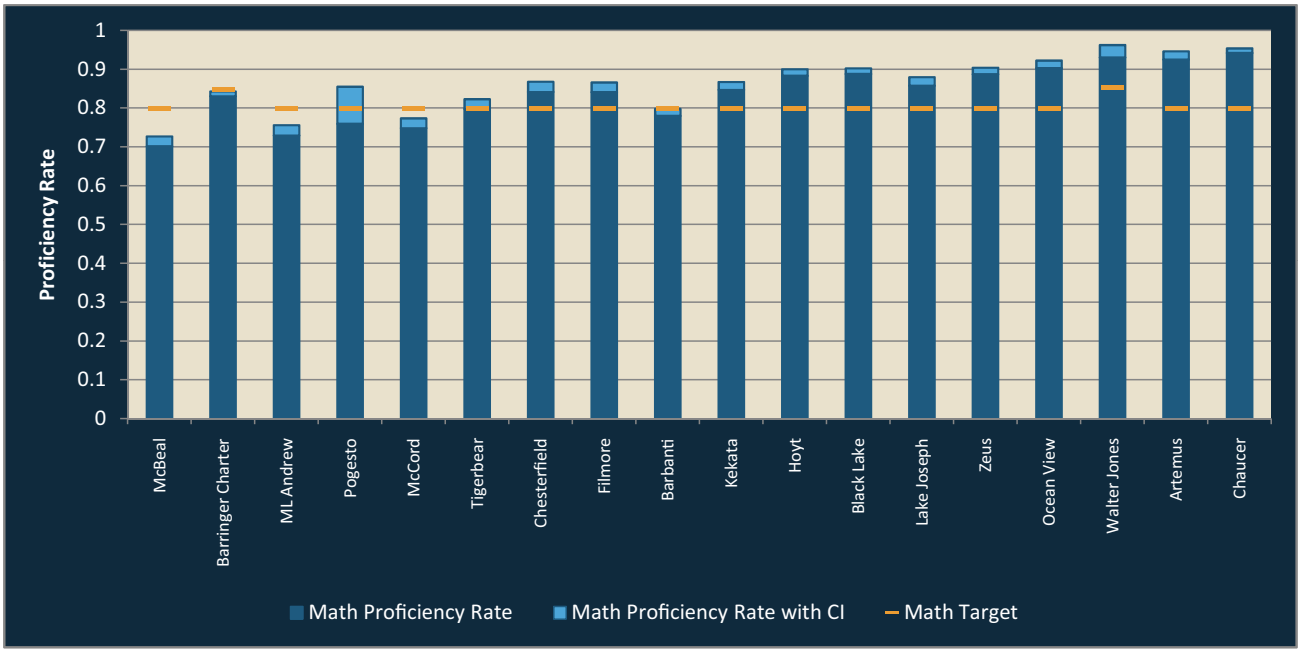

Figure 6. Impact of the confidence interval on middle school math proficiency rates under the Colorado 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools, Pogesto, was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

ance (Figure 4). **In short, applying the confidence interval has very modest impact on AYP decisions for the sample elementary and middle schools in Colorado.**[10]

Where do schools fail?

Figures 3 and 4 illustrate how schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Colorado rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Overall, most elementary schools performed fairly well in terms of meeting AYP targets.
- Three elementary schools failed to meet reading targets for their overall school population. No elementary schools failed in math.
- Four middle schools failed to meet math targets for their overall population and five failed in reading.

[10] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Colorado AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	86.2%	76.9%	Y	N	N	N	N	N	N	N					N	N					10	1	10%	N	1
Maryweather	87.2%	76.7%	Y	N	N	N	N	N	Y	N					Y	N			Y	Y	12	5	42%	N	1
Few	89.7%	80.8%	Y	N	N	N	N	N	Y	N					Y	N			Y	Y	12	5	42%	N	1
Nemo	91.2%	89.8%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	93.3%	88.1%	Y	Y				N	Y	Y					Y	N			Y	Y	9	7	78%	N	5
JFK	95.9%	88.4%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	3
Scholls	96.3%	90.3%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	7
Hissmore	94.3%	91.6%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	91.3%	89.0%	Y	Y	N	N		N	Y	Y					Y	Y			Y	Y	11	8	73%	N	5
Alice Mayberry	97.9%	93.4%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	9
Wayne Fine Arts	97.7%	98.9%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	97.2%	95.3%	Y	Y	Y	Y									Y	Y			Y	Y	8	8	100%	Y	22
Coastal	94.3%	89.7%	Y	Y	N	N	N	N	Y	N	Y	Y			Y	N			Y	Y	14	8	57%	N	3
Paramount	91.4%	90.3%	Y	Y					N	N					Y	N			Y	Y	8	5	63%	N	7
Forest Lake	98.7%	96.0%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	98.9%	96.4%	Y	Y	Y	Y			Y	Y									Y	Y	8	8	100%	Y	10
Roosevelt	99.3%	99.0%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	98.6%	97.6%	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y			Y	Y	12	12	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

- Three (Scholls, Alice Mayberry, Forest Lake) of the twelve failing elementary schools didn't make AYP because of one target.
- Every LEP subgroup and almost every SWD subgroup at the middle school level did not meet targets in reading and math.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively.[11] As shown, the performance of students with disabilities is proving most challenging for schools under Colorado's system, particularly for middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 30. In fact, every single middle school with a SWD population large enough to qualify as a separate subgroup failed to meet its math and reading targets for these students (except Ocean View). Students with LEP also struggled to meet the state's targets; all middle schools with a LEP population large enough to qualify

[11] Recall that elementary students do better on Colorado's math test than middle school students perhaps because Colorado's proficiency scores are easier in math than in reading at the elementary grades (see Figure 2).

Table 3. Middle school subgroup performance of sample schools under the 2008 Colorado AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	70.2%	75.1%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	83.0%	85.4%	N	N	N	N			N	N	N	N			Y	Y					10	2	20%	N	0
ML Andrew	72.9%	83.3%	N	N	N	N			N	N	N	N			Y	N			Y	Y	12	3	25%	N	0
Pogesto	75.9%	88.9%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	74.8%	85.2%	N	Y	N	N			N	N	N	N			N	Y			Y	Y	12	4	33%	N	0
Tigerbear	79.7%	81.4%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	0
Chesterfield	84.1%	84.8%	Y	Y	N	N			Y	N	N	N							Y	Y	10	5	50%	N	1
Filmore	84.1%	89.4%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	1
Barbanti	78.0%	83.5%	Y	N	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	84.7%	85.3%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	0
Hoyt	88.3%	88.7%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	2
Black Lake	88.8%	88.6%	Y	Y	N	N	N		Y	N	Y	N	Y	Y	Y	N			Y	Y	15	9	60%	N	0
Lake Joseph	85.8%	90.0%	Y	Y	N	N	N	N	Y	Y					N	N			Y	Y	12	6	50%	N	2
Zeus	88.7%	88.6%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	1
Ocean View	90.3%	94.1%	Y	Y	N	Y	N	N	N	Y					N	Y			Y	Y	12	7	58%	N	2
Walter Jones	93.0%	93.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	92.5%	90.9%	Y	Y	N	N			N	N			Y	Y	N	N			Y	Y	12	6	50%	N	3
Chaucer	94.2%	96.1%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP

as a separate subgroup failed to meet math and reading targets for these students.

Moreover, Hispanic students in Colorado struggled to meet targets as well. At the elementary level, 6 of the 9 qualifying subgroups failed to meet their reading targets. At the middle school level, 6 of 14 qualifying subgroups failed to meet both reading and math targets.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Colorado's NCLB accountability system is, in most respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Colorado, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Colorado.

One exception is Nemo elementary school (see Figure 3) which failed to make AYP in 21 states, yet succeeded in Colorado. Examining Table 2, we can see that Nemo didn't meet the minimum numbers for the LEP and

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Colorado AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	13	6	10
Students with limited English proficiency	7	4	6
Low-income students	17	2	5
African-American students	6	0	1
Asian/Pacific Islander students	0	0	0
Hispanic students	9	1	6
American Indian/Alaska Native students	0	0	0
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Colorado AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	15
Students with limited English proficiency	9	9	8
Low-income students	17	8	12
African-American students	11	6	10
Asian/Pacific Islander students	4	0	0
Hispanic students	14	8	8
American Indian/Alaska Native students	1	1	1
White students	17	0	0

SWD subgroups, which created difficulty for many other schools in the sample. Nemo also enrolled fewer than the minimum numbers of African American or Hispanic students to qualify as accountable subgroups. With fewer subgroups, and in a state with relatively easy proficiency standards (Figure 2), Nemo made AYP in Colorado, even when other schools with higher average performance failed.

This is consistent with the patterns shown in Table 6, which compares the sample schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, elementary schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower per-

Table 6. Comparisons between schools that did and didn't make AYP in Colorado, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	6	12	2	16
Average student body size	231	342	124	951
Average % low income	19	60	42	45
Average % nonwhite	26	48	27	46
Average performance†	4.93	-0.63	0.40	-0.11
Average % growth‡	116	115	109	97
Average number of targets to meet	8	10	6	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

centages of nonwhite students. Similarly, middle schools that make AYP have slightly higher performing students, on average, than middle schools that don't make it, but have smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Colorado's AYP rules (and AMOs) for 2008. We found that only 6 elementary schools and 2 middle schools — 8 in all, from a sample of 36—would have made AYP in Colorado. Looking across the 28 state accountability systems examined in the study, this puts Colorado in the upper middle of the distribution in terms of the number of schools making AYP (see Figure 1). Colorado's cut scores are low but its annual targets for proficiency are fairly high; thus, large numbers of schools did not make AYP in Colorado despite its low proficiency standards.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Colorado's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income students. Almost all of the sample schools met the Colorado reading and math targets for their overall populations, i.e., without considering subgroup results. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and students with disabilities are more generous under the current state assessment sys-

tem,[12] doesn't the massive failure of these students, especially in middle schools, to meet Colorado's targets indicate that a new approach is needed for holding schools accountable for their performance? Yes, schools should redouble their efforts to boost achievement for LEP students and students with disabilities, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[12] See Footnote 5.

Delaware

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Delaware's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP—or not making AYP. It also gauges how tough Delaware's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Delaware's system as well as under the systems 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Delaware's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **13 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Delaware's accountability system. (This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, **we find that the number of elementary schools making AYP in Delaware was exceeded in 11 other sample states, putting Delaware roughly in the middle of the sample distribution (see Figure 1).**[2]
- Nearly all the schools in our sample that failed to make AYP in Delaware are meeting expected targets for their overall populations but failed to make AYP because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[3]
- One sample school (Alice Mayberry) that failed to make AYP in most other states made AYP in Delaware.

Looking across the 28 state accountability systems examined in the study, we find **Delaware** near the middle of the distribution in terms of how many sample schools make AYP. Delaware's mix of rules means that several schools make AYP in Delaware that do not in most of the other 27 states. This is likely due to the fact that Delaware's proficiency standards (or cut scores) are relatively easy compared to other states. However, Delaware's annual targets (i.e., the percentage of students in various subgroups who have to meet proficiency) in reading are relatively difficult to achieve. Specifically, 68 percent of a given population in any school would have to be proficient on the state reading exam for the school to make AYP in 2008. Every single school with a limited English proficient (LEP) subgroup failed to make AYP in Delaware, in part because these students did not meet the state's proficiency targets in reading or/math.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Delaware Student Testing Program.

[2] Note that Delaware received full approval from the U.S. Department of Education to implement a student growth model for the 2006–2007 school year. The current analysis, which draws on data from 2005–2006, does not in any way use or incorporate student growth model calculations.

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

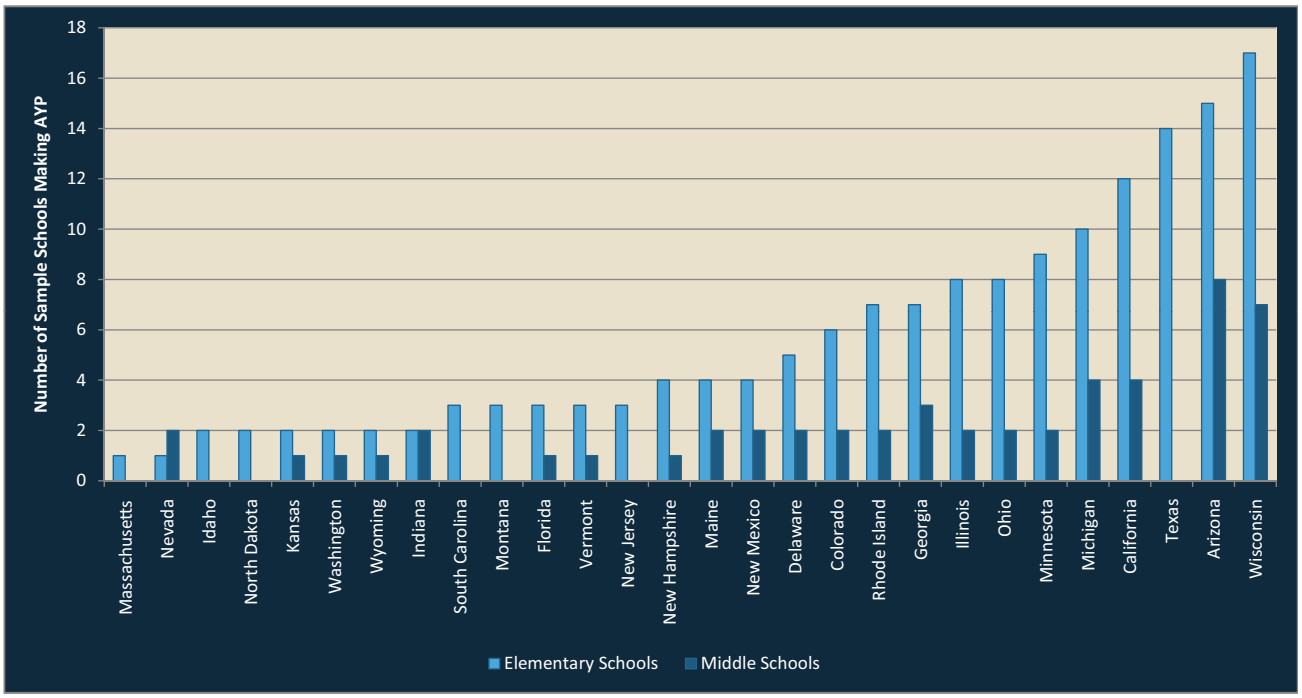

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

This is probably because **Delaware's proficiency standards are relatively easy compared to other states.**

- In Delaware, as in most states, schools with fewer subgroups attain AYP more easily in Delaware than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.
- As in other states, middle schools have greater difficulty reaching AYP in Delaware than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.
- A strong predictor of a school making AYP under Delaware's system is whether it has enough English language learners to qualify as a separate subgroup. Every school with a subgroup of students with limited English proficiency (LEP)[4] failed to make AYP, in part because these students did not meet the state's proficiency targets in reading and/or math. Likewise, many schools with enough qualifying students with disabilities (SWDs) failed to meet their AYP targets.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Delaware's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state compar-

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

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Delaware Student Testing Program (DSTP), the standardized state test. 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.

isons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the Delaware Student Testing Program (DSTP) are taken from *The Proficiency*

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

[7] We gave all schools in our sample pseudonyms in this report.

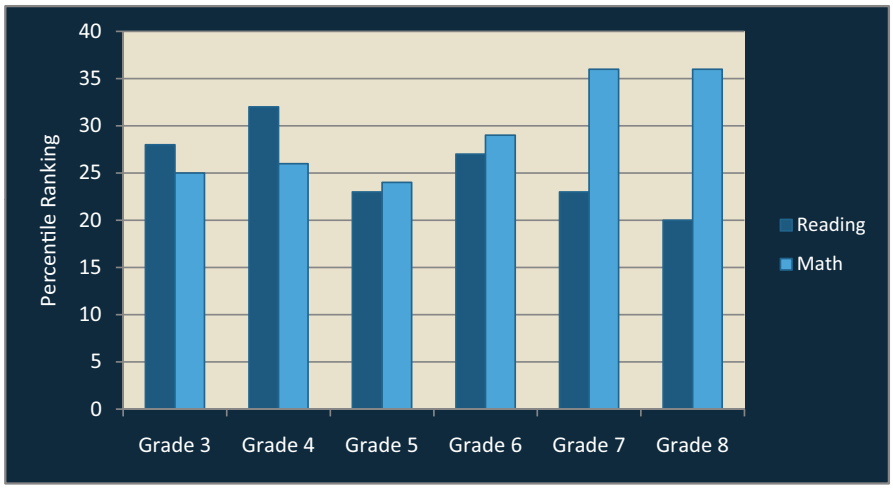

Figure 2. Delaware reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Delaware's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Delaware's cut scores are below the 40th percentile.

Illusion (as shown in Figure 2), which found that Delaware's definitions of proficiency generally ranked below the average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Delaware test, given their performance on MAP. Student test data and subgroup designations are then used to determine how these 18 elementary and 18 middle schools would have fared under Delaware AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Delaware's 2008 AYP rules.

Table 1 shows the pertinent Delaware AYP rules that were applied to elementary and middle schools in this study. Delaware's minimum subgroup size is 40, which is comparable to most other states we examined.[8] Furthermore, although most states examined in the study apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates, **Delaware's 98% confidence interval gives schools greater leniency than the 95% confidence interval used by most other states**. So, for instance, though schools are supposed to get 68% of their students (as well as 68% of their students in each subgroup) to the proficient level on the state reading test, applying the confidence interval means that the real target can actually be lower, particularly with smaller groups.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students, and 95% of the students in each school's subgroup, to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence inter-

[8] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

Table 1. Delaware AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 40	
	Low-income students: 40	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 98% CI	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	62	68
Grade 4	62	68
Grade 5	62	68
Grade 6	62	68
Grade 7	62	68
Grade 8	62	68
MATH		
Grade 3	41	50
Grade 4	41	50
Grade 5	41	50
Grade 6	41	50
Grade 7	41	50
Grade 8	41	50

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

vals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Delaware's AYP Rules?

Figure 3 illustrates the AYP performance of the sample **elementary schools** under Delaware's 2008 AYP rules. **Only 5 schools made AYP and 13 failed to make AYP.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student and positive values indicating above-grade-level performance. All schools that made AYP are in the right half of the figure, meaning that the higher performing students were found at these schools.

Yet almost without regard to average student performance, the only schools actually to make AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has separate targets). For example, Wayne Fine Arts and Winchester passed, but had only four targets each. Each school must make AYP for its overall student population in reading and math (two targets) and for its white population resulting in four total targets.

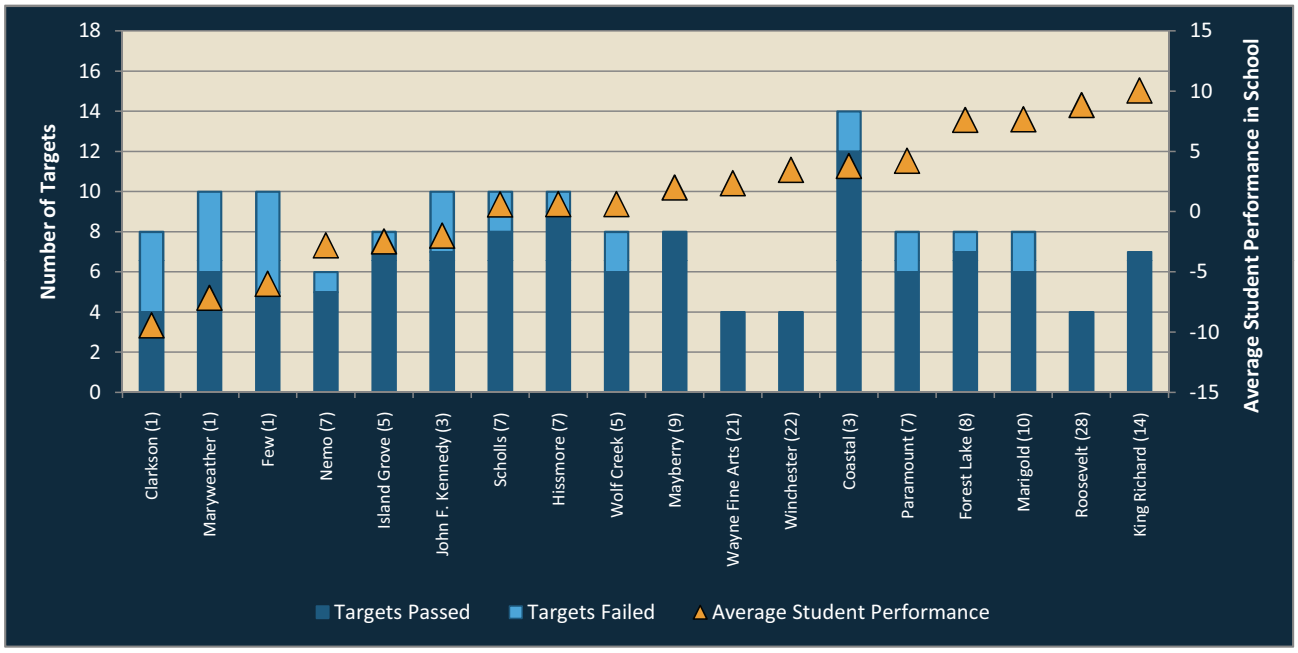

Figure 3. AYP performance of the elementary school sample under Delaware's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Delaware's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Wolf Creek Elementary, for example, meets six of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Delaware AYP rules. **Out of 18 middle schools in our sample, only 2 passed**—one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Figure 5 indicates the degree to which elementary schools' math proficiency rates are aided by the confidence interval. On this figure, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates were increased by applying the confidence interval. The orange lines show the annual measurable objective needed to meet AYP. The figure shows that none of the sample elementary schools was assisted by the confidence intervals, because the annual mathematics targets in Delaware are already low (i.e., 50%, see Table 1) relative to schools' overall performance. The effect of confidence intervals on middle school math proficiency rates and the reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, none of the sample elementary or middle schools is assisted by the confidence intervals. **In short, applying the confidence interval (even a generous one like the 98% confidence interval used in Delaware) has little or no effect on whether schools meet their overall reading and math targets in Delaware, mostly because of the state's low annual targets.**[9]

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figure 5. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

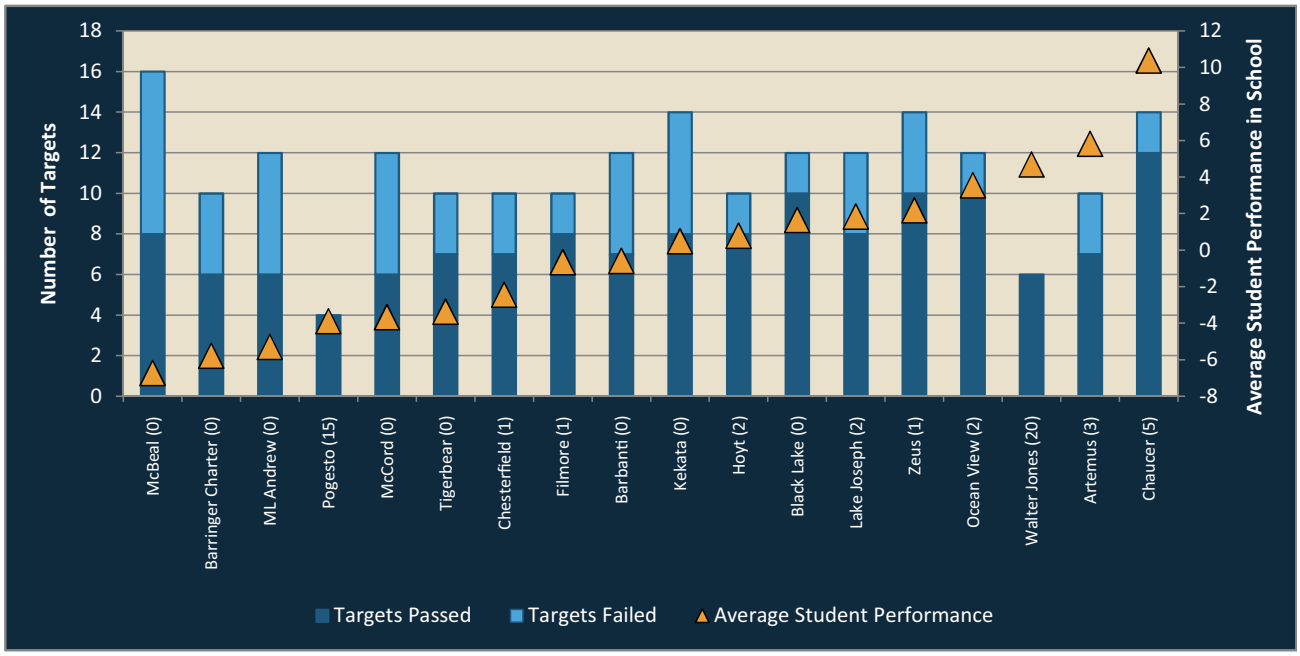

Figure 4. AYP performance of the middle school sample under Delaware's 2008 AYP rules

Note: This figure shows how each middle school within the sample would have fared under Delaware's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means that the school failed. Artemus Middle School, for example, met 7 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

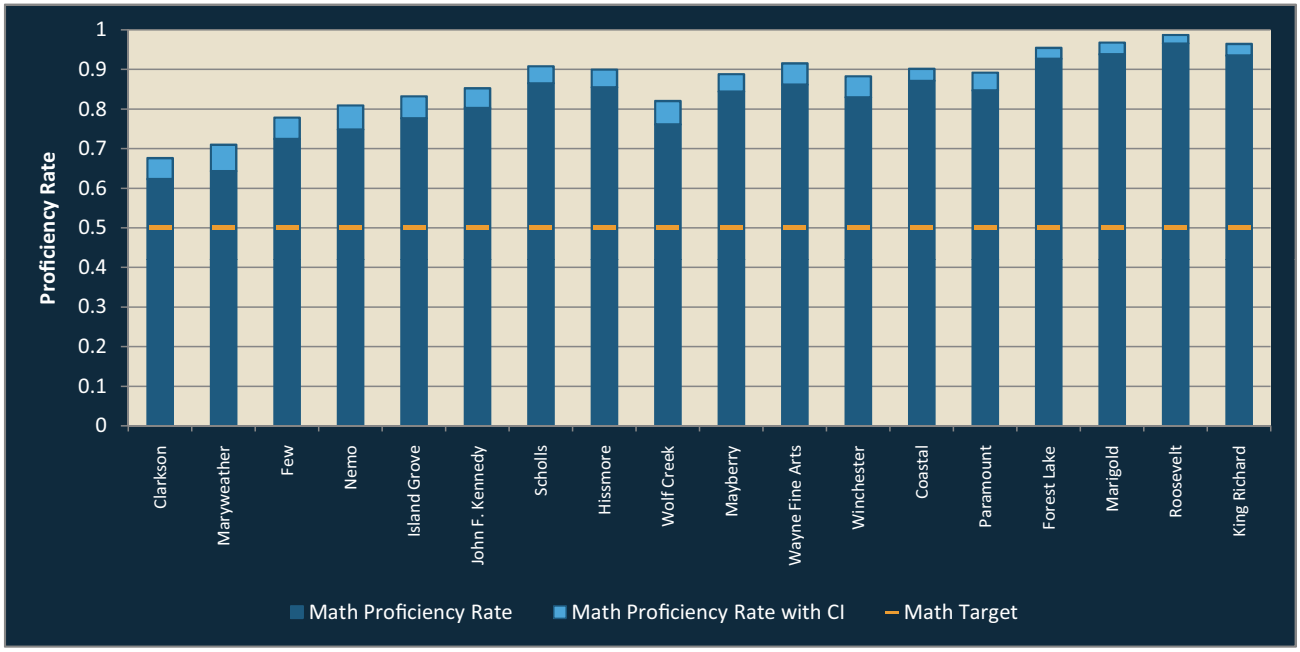

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Delaware AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	62.4%	47.3%	Y	N			Y	N	Y	N					Y	N					8	4	50%	N	1
Maryweather	64.4%	53.4%	Y	N			Y	N	Y	N					Y	N			Y	Y	10	6	60%	N	1
Few	72.5%	59.1%	Y	N	Y	N	Y	N	Y	N					Y	N					10	5	50%	N	1
Nemo	74.9%	71.2%	Y	Y					Y	N									Y	Y	6	5	83%	N	7
Island Grove	77.7%	70.4%	Y	Y					Y	Y					Y	N			Y	Y	8	7	88%	N	4
JFK	80.3%	66.8%	Y	Y	Y	N			Y	N	Y	N							Y	Y	10	7	70%	N	3
Scholls	86.6%	72.1%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	7
Hissmore	85.6%	75.2%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	7
Wolf Creek	76.1%	72.1%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	5
Alice Mayberry	84.5%	79.2%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	86.2%	85.6%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	83.0%	82.9%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	87.2%	78.2%	Y	Y	Y	N	Y	N	Y	Y	Y	Y			Y	Y			Y	Y	14	12	86%	N	3
Paramount	84.8%	78.4%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	92.8%	87.4%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	93.9%	88.1%	Y	Y	Y	N			Y	N									Y	Y	8	6	75%	N	10
Roosevelt	96.6%	93.9%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	93.6%	91.2%	Y	Y	Y	Y			Y										Y	Y	7	7	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Delaware rules, and the total number of states within the study in which that school met AYP. The school-by-school findings in Tables 2 and 3 show that:

- Three elementary schools (Clarkson, Maryweather, and Few) failed to meet reading targets for their

Table 3. Middle school subgroup performance of sample schools under the 2008 Delaware AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	57.5%	65.2%	Y	Y	N	N	N	N	N	N	Y	Y			N	N	Y	Y	Y	Y	16	8	50%	N	0
Barringer Charter	63.2%	66.6%	Y	Y	N	N			Y	N	Y	N			Y	Y					10	6	60%	N	0
ML Andrew	55.8%	71.9%	Y	Y	N	N			N	N	N	N			Y	Y			Y	Y	12	6	50%	N	0
Pogesto	53.7%	77.8%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	58.6%	73.3%	Y	Y	N	N			N	N	N	N			Y	Y			Y	Y	12	6	50%	N	0
Tigerbear	67.2%	69.7%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	0
Chesterfield	70.7%	73.6%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	1
Filmore	71.2%	80.2%	Y	Y	N	N			Y	Y					Y	Y			Y	Y	10	8	80%	N	1
Barbanti	65.2%	75.6%	Y	Y	N	N	N	N	Y	N					Y	Y			Y	Y	12	7	58%	N	0
Kekata	73.3%	76.8%	Y	Y	N	N	N	N	Y	Y	Y	N			Y	N			Y	Y	14	8	57%	N	0
Hoyt	76.8%	80.4%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	2
Black Lake	79.5%	81.0%	Y	Y	N	N			Y	Y	Y	Y			Y	Y			Y	Y	12	10	83%	N	0
Lake Joseph	75.1%	84.9%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	79.0%	81.7%	Y	Y	Y	N	N	N	Y	Y	Y	Y			Y	N			Y	Y	14	10	71%	N	1
Ocean View	81.5%	89.1%	Y	Y	Y	Y	N	N	Y	Y					Y	Y			Y	Y	12	10	83%	N	2
Walter Jones	85.5%	86.3%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	85.0%	85.1%	Y	Y	Y	N			Y	N					Y	N			Y	Y	10	7	70%	N	3
Chaucer	87.4%	92.6%	Y	Y	N	Y	Y	N	Y	Y			Y	Y	Y	Y			Y	Y	14	12	86%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

overall school population.

- All elementary schools met math targets for their overall population, as did all middle schools for both reading and math.
- Two of the 13 elementary schools (Hissmore and Forest Lake) and 3 of the 16 middle schools (Filmore, Hoyt, and Black Lake) that didn't make AYP only for their SWDs.
- One elementary school (Nemo) failed to make AYP only because of its low-income subgroup, and one elementary school (Island Grove) passed in every subgroup except for Hispanic students.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively.[10] As shown, the performance of students with disabilities is proving most challenging for schools under Delaware's system,

[10] Recall that elementary students do better on Delaware's math test than middle school students, perhaps because Delaware's cut scores are lower in math than in reading in grades 3 and 4 (see Figure 2).

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Delaware AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	0	7
Students with limited English proficiency	4	0	4
Low-income students	15	0	8
African-American students	5	0	2
Asian/Pacific Islander students	0	0	0
Hispanic students	7	0	6
American Indian/Alaska Native students	0	0	0
White students	16	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Delaware AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	13	14
Students with limited English proficiency	7	6	7
Low-income students	17	3	6
African-American students	10	2	6
Asian/Pacific Islander students	1	0	0
Hispanic students	13	1	4
American Indian/Alaska Native students	1	0	0
White students	17	0	0

particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum n of 40. In fact, all but one elementary school in the study with qualifying SWD subgroups failed to make AYP. Students with LEP are also struggling to meet the state's targets; every school with a large enough LEP population to qualify as a separate subgroup failed to meet its reading targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 2 and 3 indicates that Delaware's NCLB accountability system is, in most respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest

Table 6. Comparisons between schools that did and didn't make AYP in Delaware, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	5	13	2	16
Average student body size	265	320	124	951
Average % low income	24	55	42	45
Average % nonwhite	30	45	27	46
Average performance†	5.35	-0.36	0.40	-0.11
Average % growth‡	113	115	109	97
Average number of targets to meet	5	9	5	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

number of states—28, 22, and 21, respectively. And these schools all made AYP in Delaware, too. Likewise, the elementary and middle schools that fail to make AYP in the greatest number of states also failed to make AYP in Delaware.

But Delaware is also home to a few anomalies. First, consider Mayberry Elementary (see Figure 3). It failed to make AYP in 19 of the 28 states in our sample, yet made AYP in Delaware. In examining Table 2, we can see that Mayberry didn't meet the minimum numbers for the students with LEP or SWD subgroups, which create difficulty for so many other schools in the study. With fewer accountable subgroups and relatively easy proficiency standards (Figure 2), Mayberry made AYP even when other schools with higher average performance didn't. Second, look at Pogesto Middle School (Figure 4). Even with its relatively low average performance, it made AYP in Delaware, but failed to do so in 13 of 28 states. Like Mayberry, its AYP success in Delaware is most likely attributable to its relatively small number of targets (four) along with Delaware's relatively easy proficiency standards compared to other states.

This is consistent with the patterns shown in Table 6, which compares schools making and not making AYP on a number of academic and demographic dimensions. Within the sample, elementary schools that made AYP did indeed show higher average student performance, but they also differed in the following ways: they had smaller student populations, fewer subgroups (and thus fewer targets to meet), and lower percentages of low-income and minority students. Similarly, middle schools that made AYP had slightly higher performing students, on average, than middle schools that failed, but they also had dramatically smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

The study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Delaware's AYP rules (and AMOs) for 2008. We found that only 5 elementary schools and 2 middle schools—7 in all, from a sample of 36—would have made AYP in Delaware. Looking across the 28 state accountability sys-

tems examined in the study, this puts Delaware roughly in the middle of the sample distribution, as shown in Figure 1. In addition, Delaware uses a generous 98% confidence interval, but it appears to have little or no effect on whether schools meet their overall reading and math targets because the state already has such low annual targets compared to other states.

The overriding goal of the federal NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Delaware's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students such as low-income or Hispanic students. Almost all the sample schools made AYP in Delaware for their student populations as a whole (i.e., without considering subgroup results). In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[11] doesn't the failure of these students to meet Delaware's targets (especially at the middle school level) indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do

[11] See footnote 5.

not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Florida

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Florida's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Florida's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether or not each would make AYP under Florida's system as well as in systems in 27 other states. We used school data estimates from academic year 2005–2006, but applied them against Florida's AYP rules and cut scores[1] for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **15 of 18 elementary schools** and **17 of 18 middle schools** in our sample failed to **make AYP** in 2008 under Florida's accountability system. (The high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, only 8 states passed fewer of the sample elementary schools than Florida, while 4 states tied with Florida. **In addition, Florida was one of 6 states with a single middle school that made AYP in the sample (see Figure 1).**[2]
- Many of the schools in our sample that failed to make AYP in Florida met expected targets for their overall populations but didn't make AYP because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[3]
- **Two sample schools that failed to make AYP in most other states made AYP in Florida. This is**

Only four schools in the study make AYP in **Florida**. This can be attributed to a couple of factors. First, Florida's cut scores range from the 30th to the 53rd percentile; hence, proficiency standards are relatively hard to achieve. Florida also does not apply a confidence interval (margin of error) to proficiency rate calculations (percentage of students achieving proficient or higher on the state test). This means that in Florida, schools will have greater difficulty achieving their annual targets than they would in states that employ confidence intervals. On the other hand, a couple of schools in the study make AYP in Florida but don't in most other states. This is likely because the minimum subgroup size in Florida tends to be large, meaning Florida schools will be accountable to fewer groups than schools in many other states.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Florida Comprehensive Assessment Test.

[2] Note that Florida received full approval from the U.S. Department of Education to implement a student growth model for the 2006-2007 school year. The current analysis, which draws on data from 2005–2006, does not in any way use or incorporate student growth model calculations.

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

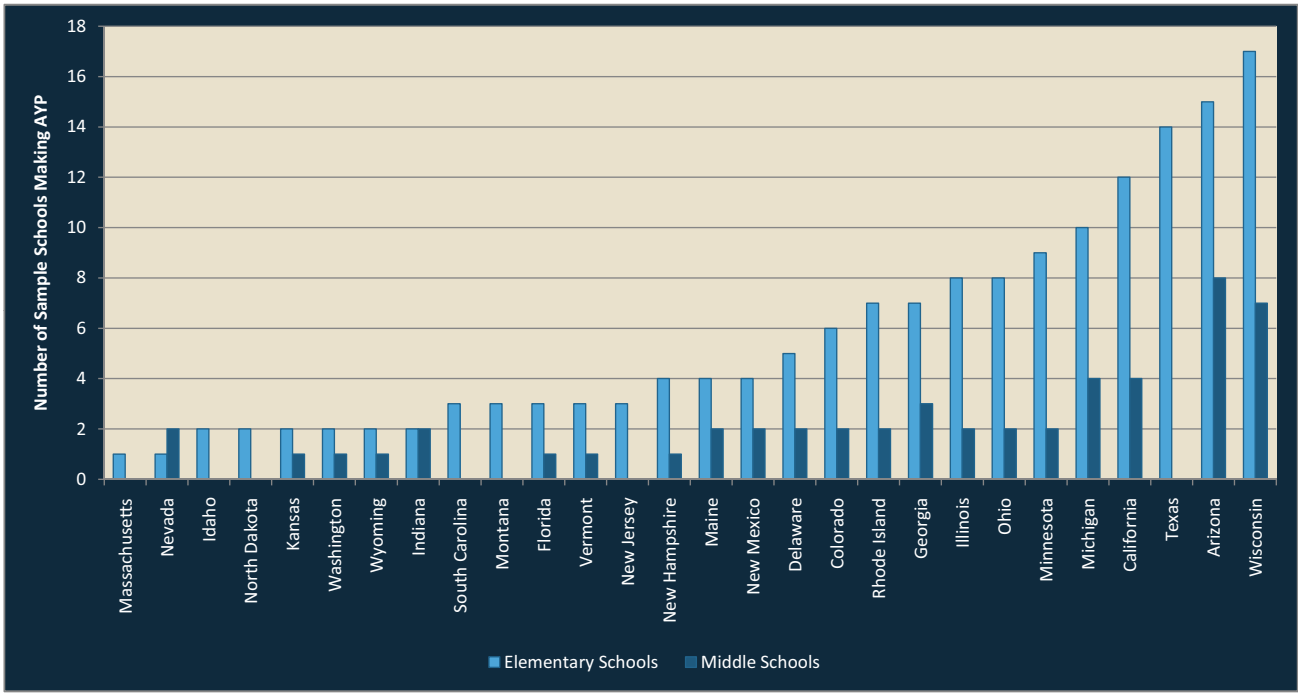

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

probably because these two schools had fewer accountable subgroups under Florida's AYP rules.

- Schools with fewer subgroups attained AYP more easily in Florida than schools with more subgroups, even when their average student performance was much lower. In other words, schools with greater diversity and size face greater challenges in making AYP. This is true other states as well.

- Middle schools had greater difficulty reaching AYP in Florida than did elementary schools, primarily because some of the middle school proficiency standards are more difficult than at the elementary grades, and because the student populations are larger and therefore the schools have more qualifying subgroups—not because their student achievement was lower than in the elementary schools.

- A strong predictor of whether or not a school would make AYP under Florida's system is whether it has enough English language learners or SWDs to qualify as separate subgroups. Every school with a limited English proficient (LEP)[4] or SWD subgroup failed to make AYP.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on various standardized tests in 25 states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed

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

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, mostly because the likely differences between how LEP students and SWDs are treated in MAP, the assessment we used in this study, and in the Florida Comprehensive Assessment Test, the standardized state test. Specifically, the U.S. Department of Education has issued NCLB guidelines in recent years that exclude small percentages of LEP students and SWDs from taking the state test or allow them to take alternative assessments. In this study, however, no valid MAP scores were omitted from consideration.

profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* (for which Florida did not participate) by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus Florida, and 2 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low-income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

These schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the last calculated by the percentage of students receiving free or reduced price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the Florida Comprehensive Assessment Test (FCAT) are shown in Figure 2. These cut scores were used to estimate whether students would have scored as proficient or better on the Florida

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

[7] We gave all schools in our sample pseudonyms in this report.

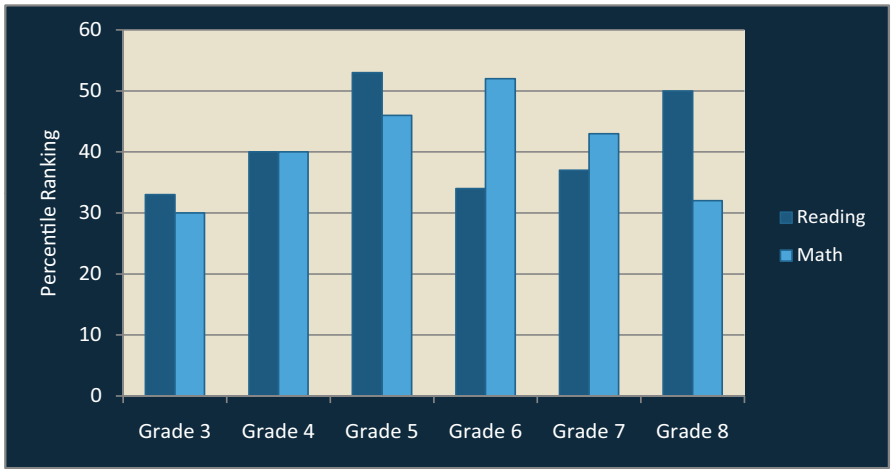

Figure 2. Florida reading and math cut score estimates, expressed as percentile ranks (2007)

Note: This figure illustrates the difficulty of Florida's cut scores (proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm groups, in grades three through eight. Percentile ranks denote the percentage of the NWEA norm group that would perform at or below that standard. For example, 70% of the third graders in NWEA's norm group would have exceeded the performance necessary to achieve math proficiency on the FCAT. Higher percentile ranks are more difficult to achieve. Most of Florida's cut scores are near or below the 50th percentile.

Table 1. Florida AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30 or 15% of school population, up to 100 students	
	SWDs: 30 or 15% of school population, up to 100 students	
	Low-income students: 30 or 15% of school population, up to 100 students	
	LEP students: 30 or 15% of school population, up to 100 students	
CI	**Applied to proficiency rate calculations?**	
	Not used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	31	58
Grade 4	31	58
Grade 5	31	58
Grade 6	31	58
Grade 7	31	58
Grade 8	31	58
MATH		
Grade 3	38	62
Grade 4	38	62
Grade 5	38	62
Grade 6	38	62
Grade 7	38	62
Grade 8	38	62

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).
Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

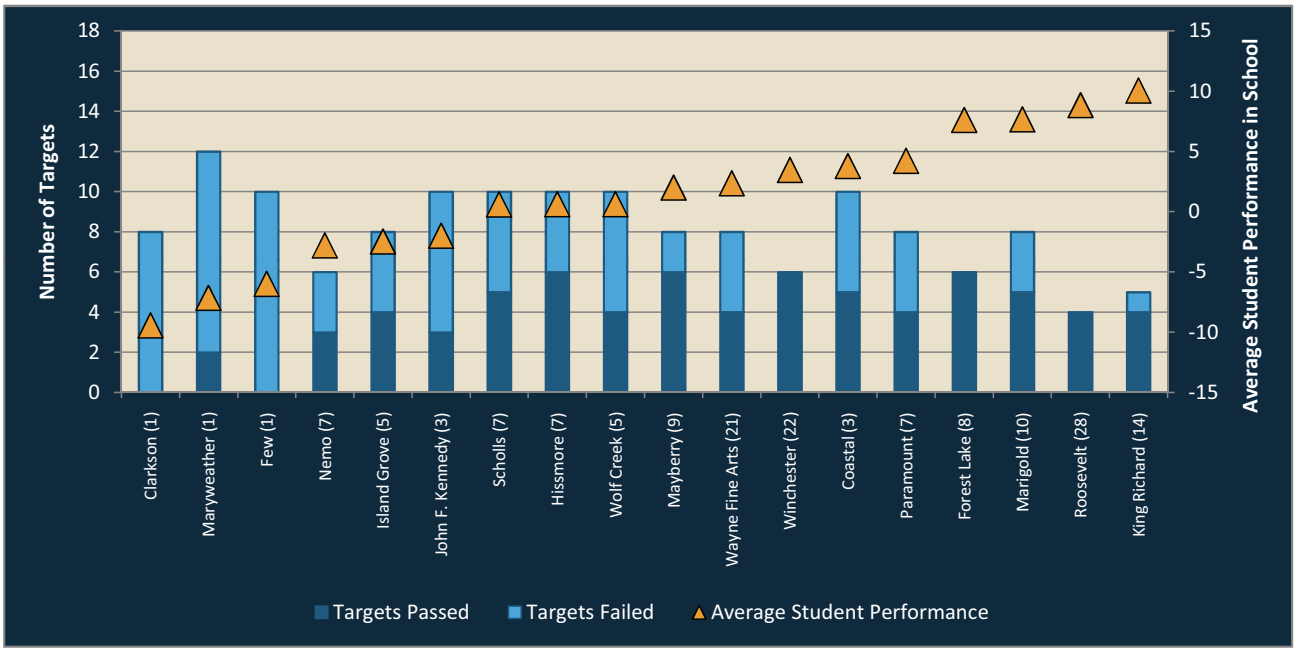

Figure 3. AYP performance of the elementary school sample under Florida's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Florida's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. King Richard, for example, met four of its five targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; this scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Florida AYP rules for 2008. (In other words, the school data are from academic year 2005–2006, but we are applying them against Florida's 2007–2008 cut scores and AYP rules.)

Table 1 shows the pertinent Florida AYP rules that were applied to elementary and middle schools in this study. Florida's minimum subgroup size is 30; if 30 does not constitute 15% of the total student population, then the minimum *n* is 15% of the total student population, up to 100 students.[8] **This means that for many schools the actual subgroup size is much larger than 30, meaning that Florida schools will have fewer subgroups for which its held accountable than do schools in many other states.**[9]

Unlike most other states examined in the current study, Florida does not apply confidence intervals (or margins of error) to its measurements of student proficiency rates. This means that in Florida, **schools will have greater difficulty achieving their annual measurable objectives than they would in states that employ confidence intervals.**

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% rela-

[8] So then, the minimum subgroup size in Florida cannot be less than 30 or more than 100 students. For example, a school with a total population of 1000 would have a minimum subgroup size of 100 since 30 does not constitute 15% and 15% of 1000 (i.e., 150) exceeds the 100-student ceiling.

9 Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

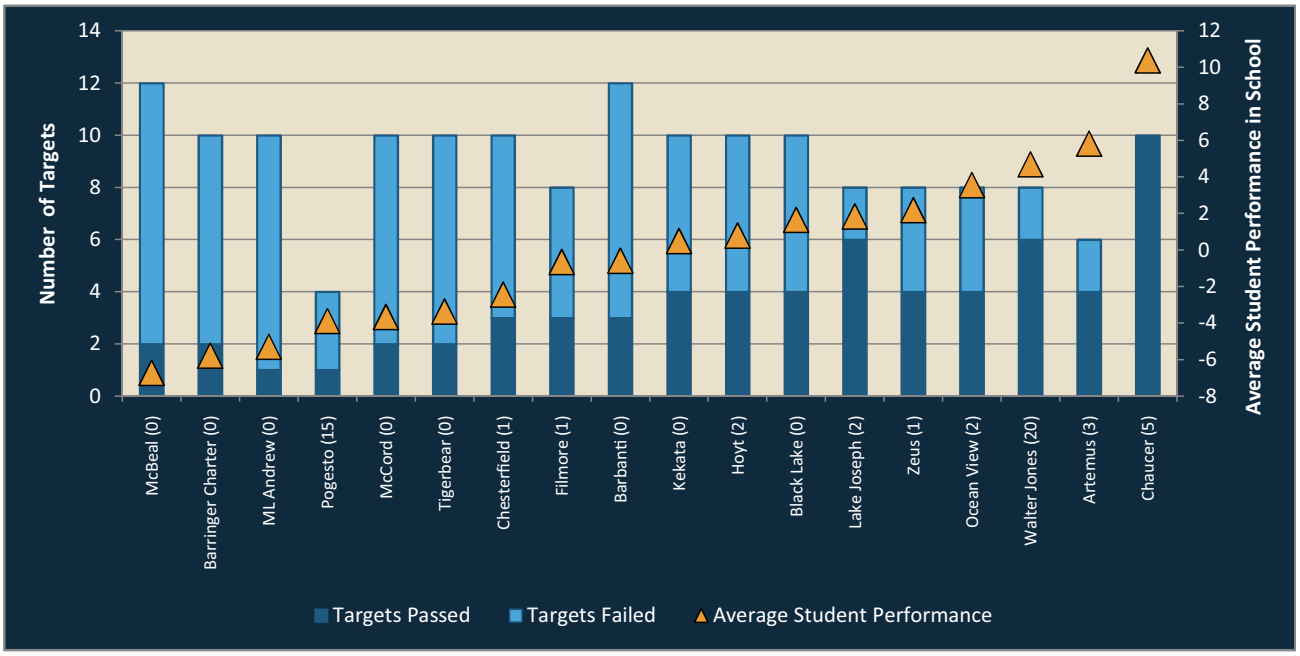

Figure 4. AYP performance of the middle school sample under Florida's 2008 AYP rules

Note: This figure shows how each middle school within the sample fared under Florida's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that fails to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Lake Joseph, for example, met six of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

tive to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Florida's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Florida's 2008 AYP rules. **Only 3 elementary schools (Winchester, Forest Lake, and Roosevelt) out of 18 made AYP**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student and positive values indicating above-grade-level performance. All schools that made AYP are in the right half of the figure, meaning that the students with the highest average performance were found at these schools.

Yet almost without regard to average student perform-

Table 2. Elementary school subgroup performance of sample schools under the 2008 Florida AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	44.9%	27.9%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	53.0%	41.6%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	1
Few	58.4%	42.3%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Nemo	60.9%	58.1%	N	Y					N	N									Y	Y	6	3	50%	N	7
Island Grove	64.3%	59.3%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	4
JFK	67.7%	51.9%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	3
Scholls	75.0%	59.9%	Y	Y	N	N			Y	N	N	N							Y	Y	10	5	50%	N	7
Hissmore	75.7%	61.8%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	7
Wolf Creek	67.0%	61.2%	Y	Y	N	N			N	N					N	N			Y	Y	10	4	40%	N	5
Alice Mayberry	73.1%	61.9%	Y	Y					Y	N	Y	N							Y	Y	8	6	75%	N	9
Wayne Fine Arts	71.8%	71.3%	Y	Y					N	N	N	N							Y	Y	8	4	50%	N	21
Winchester	74.5%	71.1%	Y	Y											Y	Y			Y	Y	6	6	100%	Y	22
Coastal	76.9%	63.9%	Y	Y	N	N			Y	N	N	N							Y	Y	10	5	50%	N	3
Paramount	78.1%	68.7%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	7
Forest Lake	86.6%	78.8%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	8
Marigold	88.5%	76.9%	Y	Y	N	N			Y	N									Y	Y	8	5	63%	N	10
Roosevelt	90.9%	85.4%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	86.5%	82.7%	Y	Y	N														Y	Y	5	4	80%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

ance, the only schools that made AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has its own separate targets). For example, Winchester and Forest Lake passed, but had only six targets each—two in reading and math for their overall populations, two in reading and math for their white population, and two in reading and math for an additional subgroup (Hispanic for Winchester, low income for Forest Lake).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Florida AYP rules. **Of 18 middle schools in our sample, only a single school passed**—Chaucer—the school with the highest average student performance.

Where Do Schools Fail?

Figures 3 and 4 illustrate how the elementary and middle schools, respectively, within the sample fared under the Florida rules, but do not identify which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

Table 3. Middle school subgroup performance of sample schools under the 2008 Florida AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	49.3%	48.5%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Barringer Charter	49.9%	48.4%	N	N	N	N			N	N	N	N			Y	Y					10	2	20%	N	0
ML Andrew	48.3%	51.4%	N	N					N	N	N	N			N	N			N	Y	10	1	10%	N	0
Pogesto	53.7%	53.7%	N	N															N	Y	4	1	25%	N	15
McCord Charter	48.9%	57.4%	N	N					N	N	N	N			N	N			Y	Y	10	2	20%	N	0
Tigerbear	58.4%	50.9%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	0
Chesterfield	63.7%	52.4%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	1
Filmore	60.6%	61.7%	N	Y					N	N					N	N			Y	Y	8	3	38%	N	1
Barbanti	59.2%	58.3%	N	Y	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	67.7%	60.8%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	0
Hoyt	68.8%	64.5%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	2
Black Lake	73.1%	61.4%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	0
Lake Joseph	70.2%	66.9%	Y	Y					Y	Y					N	N			Y	Y	8	6	75%	N	2
Zeus	72.4%	67.4%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	1
Ocean View	72.9%	77.2%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	2
Walter Jones	74.4%	77.7%	Y	Y					N	Y					N	Y			Y	Y	8	6	75%	N	20
Artemus	76.1%	77.7%	Y	Y					N	N									Y	Y	6	4	67%	N	3
Chaucer	82.8%	83.3%	Y	Y					Y	Y			Y	Y	Y	Y			Y	Y	10	10	100%	Y	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*) and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP are SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Florida rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Three elementary schools (Clarkson, Maryweather, and Few) failed to meet overall population targets for both reading and math. One additional school (JFK) failed to meet the overall target in reading, and one other school (Nemo) failed to meet the overall target in mathematics.
- Six middle schools (McBeal, Barringer, ML Andrew,

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Florida AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	9	9	8
Students with limited English proficiency	3	3	3
Low-income students	15	9	14
African-American students	6	4	6
Asian/Pacific Islander students	0	0	0
Hispanic students	7	6	6
American Indian/Alaska Native students	0	0	0
White students	16	0	0

Table 5. Summary of subgroup performance of sample middle schools under the Florida AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	9	9	9
Students with limited English proficiency	2	2	2
Low-income students	17	15	14
African-American students	8	8	8
Asian/Pacific Islander students	1	0	0
Hispanic students	10	8	7
American Indian/Alaska Native students	0	0	0
White students	17	2	0

Pogesto, McCord, and Tigerbear) failed to meet overall targets in both reading and math. An additional school (Chesterfield) failed its overall target in reading, and two more (Filmore and Barbanti) failed overall targets in mathematics.

- One of the 15 elementary schools that didn't make AYP (King Richard) missed only for the SWD subgroup.
- One middle school (Artemus) failed to make AYP only because of its low-income subgroup.
- One middle school (Lake Joseph) passed in every subgroup except for Hispanic students.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, the performance of SWDs proved most challenging for schools

Table 6. Comparisons between schools that did and didn't make AYP in Florida, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	3	15	1	17
Average student body size	243	317	1083	846
Average % low income	20	52	10	47
Average % nonwhite	21	45	29	45
Average performance†	6.65	0.14	10.38	-0.67
Average % growth‡	131	112	175	94
Average number of targets to meet	5	9	10	9

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

under Florida's system. In fact, every elementary and middle school in the sample with qualifying SWD subgroups failed to meet its targets for that population. Students with LEP also struggled to meet the state's targets; every school with a large enough LEP population to qualify as a separate subgroup failed to meet its reading and math targets for these students. It is also clear that students belonging to traditionally academically disadvantaged subgroups (low income, Hispanic, and African American, among others) also struggled under the strict Florida AYP rules—many elementary and middle schools within the sample for which these subgroups were accountable failed to meet AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Florida's NCLB accountability system is, in many respects, behaving like systems in other states. For example, among the elementary schools in our sample, Roosevelt and Winchester both made AYP in the greatest number of states—28 and 22, respectively. And these schools made AYP in Florida, too. Likewise, most of the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Florida.

But Florida is also home to a few anomalies. First, consider Wayne Fine Arts (see Figure 3). It made AYP in 21 of the 28 states in our sample, but failed to make AYP in Florida. In examining Table 2, we can see that Wayne Fine Arts failed for its low-income and African American populations. **The fact that it didn't make AYP in Florida but made AYP in most other states is likely because Florida schools report no confidence interval around their proficiency rates, making it more difficult to achieve their AMOs compared to states that do use confidence intervals.**

A second anomaly is Forest Lake, which didn't make AYP in 20 of 28 states, but made AYP in Florida. Table 2 shows that this school has a relatively homogeneous student body with no accountable subgroups other than its low-income and white populations. Florida's sliding minimum subgroup rule prevents Forest Lake from having to account for most traditionally disadvantaged populations.

Two middle school anomalies are seen in Table 3 as well. Walter Jones Middle School made AYP in 20 of 28 states but failed to make AYP in Florida, because of the math performance of the Hispanic and low-income populations. As with Wayne Fine Arts Elementary, this may be attributable to Florida's lack of use of confidence intervals, making it more difficult to achieve their AMOs than it is for states that do use them. On the other hand, Chaucer Middle School made AYP in Florida but failed to make AYP in 22 of the 27 other states. This is most likely attributable to the sliding minimum *n* policy in Florida, which means that Chaucer does not have to account for either its students with LEP population or its SWDs, two subgroups that present the greatest challenges in Florida.

These observations are consistent with the patterns shown in Table 6, which compares schools that make and do not make AYP on several academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet)—at least at the elementary school level—and much lower percentages of low-income students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Florida's AYP rules and AMOs for 2008. We found that only 3 elementary schools and 1 middle school— 4 in all, from a sample of 36—would have made AYP in Florida. Looking across the 28 state accountability systems examined in the study, this puts Florida roughly in the middle of the sample distribution as shown in Figure 1. In addition, Florida is 1 of 6 states with a single middle school that made AYP in the sample.

There are several other factors of note about Florida: First, it does not apply confidence intervals (or margins of error) to its measurement of student proficiency rates. This means that schools will have greater difficulty achieving their AMOs than they would in states that employ confidence intervals. Second, the manner in which the state defines minimum *n* sizes means that Florida schools will have fewer subgroups for which it is held accountable than do schools in many other states.

The overriding goal of the federal NCLB is to eliminate educational disparities within and across states; it is important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Florida's No Child Left Behind accountability system is working exactly as Congress intended: it is identifying as needing attention those schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Most of the elementary schools and about half of the sample middle schools made AYP in Florida for their student populations as a whole, that is, without considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils were not meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Georgia

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Georgia's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough the Georgia system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under the Georgia system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against the Georgia AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **11 of 18 elementary schools** and **15 of 18 middle schools** in our sample **fail to make adequate yearly progress** in 2008 under Georgia's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, we find that **eight states exceed Georgia in terms of the number of elementary schools making AYP (see Figure 1)**.
- Nearly all of the schools in our sample that fail to make AYP in Georgia are meeting expected targets for their overall populations but failing because of the performance of individual subgroups.[2]
- **Several sample schools made AYP in Georgia that failed to make AYP in most other states. This is likely due to the fact that Georgia's proficiency standards are relatively easy, compared to other states; these schools also had fewer accountable subgroups.**
- As in other states, schools with fewer subgroups attain AYP more easily in Georgia than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.
- As in other states, Georgia's middle schools have greater difficulty reaching AYP than do elementary schools, primarily because their student populations

Georgia's AYP rules create a set of circumstances which mean that several schools make AYP in Georgia that do not in most of the 27 other states. This is likely due to the fact that Georgia's proficiency standards (or cut scores) are relatively easy compared to other states (most are below the 25th percentile). On the other hand, Georgia's annual 2008 targets for reading are relatively difficult to achieve. In fact, roughly 73 percent of a given population in any school must obtain reading proficiency in order for the school to make AYP. Consequently, every single school with a limited English proficient (LEP) subgroup failed to make AYP.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Georgia Criterion-Referenced Competency Tests (CRCT).

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

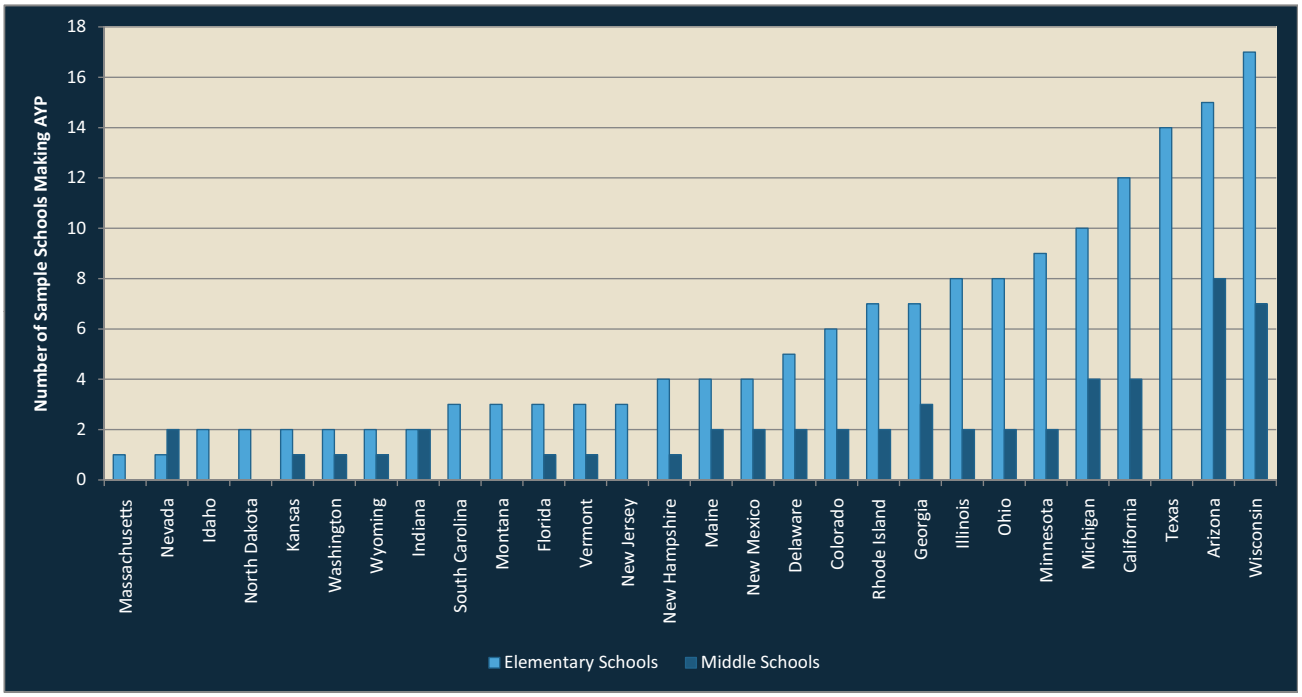

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of whether or not a school will make AYP under Georgia's system is whether it has enough students with disabilities (SWD) or English language learners to qualify as a separate subgroup. **Even though Georgia's proficiency standards (or cut scores) are relatively easy compared to other states, its annual targets (for reading, especially) are relatively difficult to achieve. Consequently, every single school with a limited English proficient (LEP)[3] subgroup failed to make AYP**, and almost all schools with enough qualifying SWD subgroups failed to meet their AYP targets.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Georgia's tests and those of 25 other states to the Northwest Evaluation Association's Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind Act (NCLB). That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Georgia Criterion-Referenced Competency Test, the standardized state test. 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.

plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low-income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target. **In Georgia, however, confidence intervals are only applied to schools that have fewer than 40 students. There were no schools that small in our sample, so confidence intervals were not considered when evaluating the performance of these schools under Georgia AYP rules.**

All of these AYP rules vary by state. This means that a school making AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Georgia Criterion-Referenced Competency Tests (CRCT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Georgia's definitions of proficiency ranked below the standards set by the other 25 states examined in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Georgia test, given their performance

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

[6] We gave all schools in our sample pseudonyms in this report.

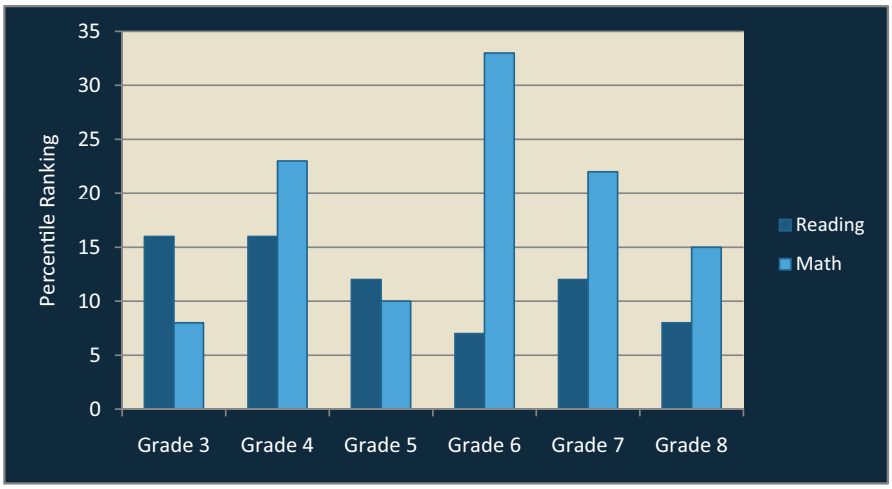

Figure 2. Georgia reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Georgia's cut scores ("proficiency passing scores") for its reading and mathematics tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Georgia's cut scores are below the 35th percentile.

Table 1. Georgia AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 40	
	Low-income students: 40	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI	Used only when school population is less than 40
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	60.0	73.3
Grade 4	60.0	73.3
Grade 5	60.0	73.3
Grade 6	60.0	73.3
Grade 7	60.0	73.3
Grade 8	60.0	73.3
MATH		
Grade 3	50.0	66.7
Grade 4	50.0	66.7
Grade 5	50.0	66.7
Grade 6	50.0	66.7
Grade 7	50.0	66.7
Grade 8	50.0	66.7

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

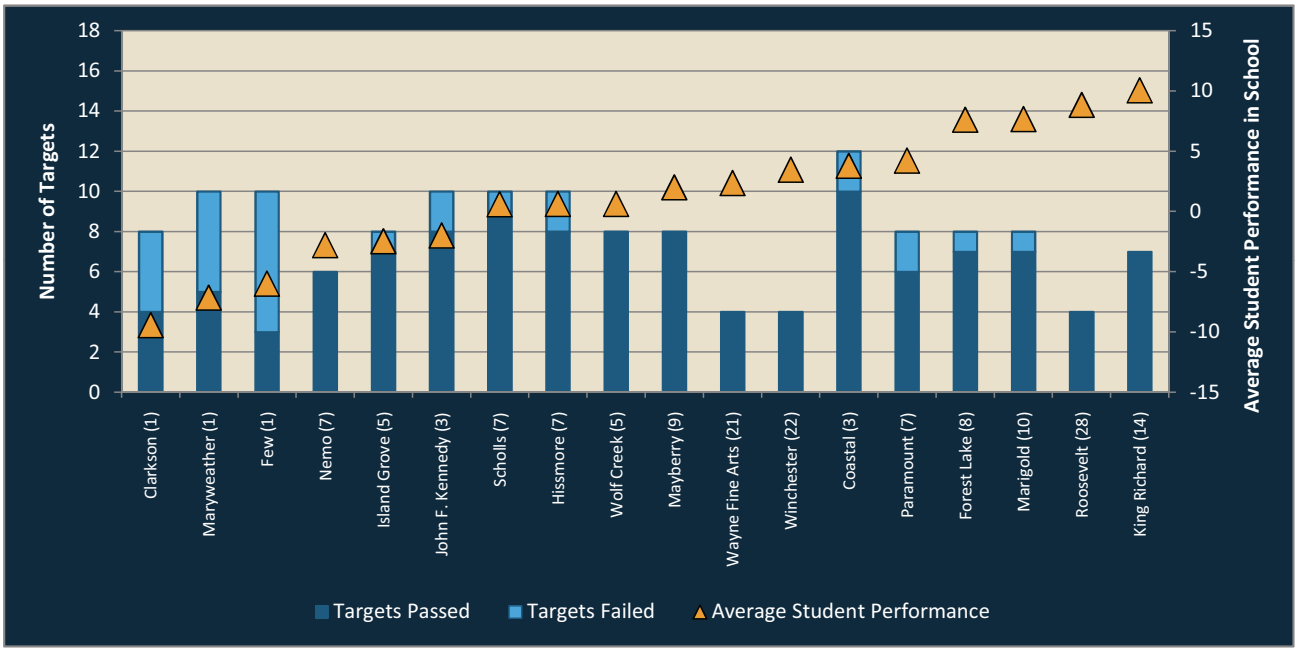

Figure 3. AYP performance of the elementary school sample under Georgia's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under the Georgia AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Paramount Elementary, for example, met six of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Georgia AYP rules for 2008. In other words, the school data are from 2005–06 but we are applying them against Georgia's 2008 AYP rules.

Table 1 shows the pertinent Georgia AYP rules that were applied to elementary and middle schools in the current study. Georgia's minimum subgroup size is 40, which is comparable to most other states in this study. Most states examined also apply confidence intervals (or margins of statistical error) to student proficiency rates. The 95% confidence interval applied by Georgia is the most commonly found confidence interval in the study. So, while schools are supposed to get 73.3% of their students to the proficient level on the state reading test, and 73.3% of their students in each subgroup, applying the confidence interval means that the real target can actually be lower, particularly with smaller groups.[7]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Most states include attendance rates as an additional indicator in their NCLB

[7] Keep in mind, however, that confidence intervals are only applied to schools in Georgia that have fewer than 40 students.

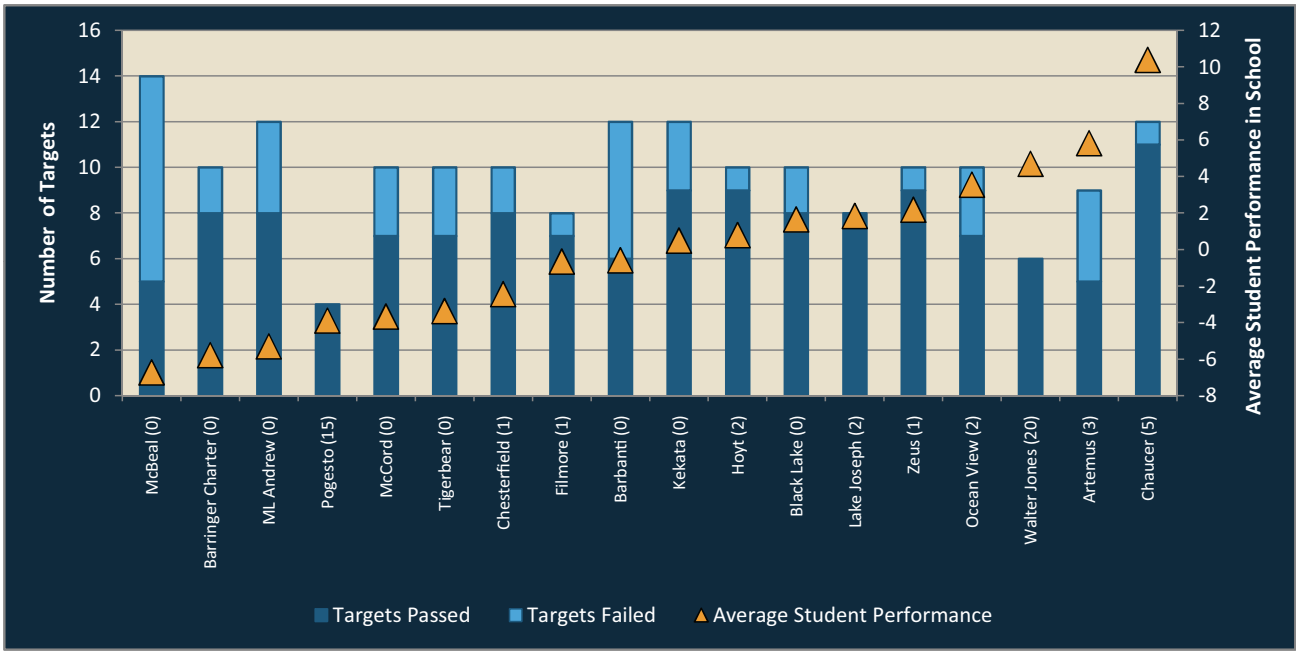

Figure 4. AYP performance of the middle school sample under Georgia's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under the AYP rules in Georgia (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them(light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Artemus Middle School, for example, met five of its nine targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

accountability system for elementary and middle schools. Plus, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under Georgia's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Georgia's 2008 AYP rules. **Only seven elementary schools made AYP while eleven failed to make it.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. One might expect that schools with higher average student performance might have better AYP outcomes than schools with lower average student performance, but as the triangles in Figure 3 show, this is not universally true. Many of the schools on the right side of the figure (higher performing) made AYP, while most of the lower performing schools did not. However, one low performing school with few subgroup targets made it (Nemo), while four higher performing schools (Coastal, Paramount, Forest Lake, and Marigold) with higher numbers of subgroup targets did not.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Georgia AYP rules. **Out of 18 in our sample, only 3 made AYP**—1 low-perfor-

Table 2. Elementary subgroup performance of sample schools under the 2008 Georgia AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	77.8%	65.6%	Y	N			Y	N	Y	N					Y	N					8	4	50%	N	1
Maryweather	78.1%	70.3%	Y	N			N	N	Y	N					Y	N			Y	Y	10	5	50%	N	1
Few	83.2%	72.5%	Y	N	N	N	N	N	Y	N					Y	N					10	3	30%	N	1
Nemo	85.6%	85.1%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	86.6%	84.0%	Y	Y					Y	Y					Y	N			Y	Y	8	7	88%	N	4
JFK	91.8%	81.3%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	3
Scholls	93.3%	85.5%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	7
Hissmore	90.9%	88.5%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	88.1%	84.5%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	5
Alice Mayberry	94.1%	90.7%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	96.0%	95.4%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	91.5%	91.0%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	90.2%	87.8%	Y	Y	Y	N			Y	Y	Y	Y			Y	N			Y	Y	12	10	83%	N	3
Paramount	89.6%	87.3%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	96.2%	94.6%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	95.7%	93.9%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	10
Roosevelt	99.0%	98.0%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	96.6%	96.3%	Y	Y	Y	Y			Y										Y	Y	7	7	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

mance school and 2 high-performance schools, but all three have relatively few qualifying subgroups.

Where do schools fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet, thanks to fewer subgroups. These figures do not, however, indicate which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Georgia rules,

Table 3. Middle school subgroup performance of sample schools under the 2008 Georgia AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	67.6%	81.0%	Y	Y	N	N	N	N	N	N	N	Y			N	N			Y	Y	14	5	36%	N	0
Barringer Charter	76.0%	84.8%	Y	Y	N	N			Y	Y	Y	Y			Y	Y					10	8	80%	N	0
ML Andrew	70.4%	90.0%	Y	Y	N	N			N	Y	N	Y			Y	Y			Y	Y	12	8	67%	N	0
Pogesto	77.8%	98.1%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	69.5%	91.9%	Y	Y					N	Y	N	Y			N	Y			Y	Y	10	7	70%	N	0
Tigerbear	77.8%	88.0%	Y	Y	N	N			Y	Y	N	Y							Y	Y	10	7	70%	N	0
Chesterfield	80.3%	91.9%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	1
Filmore	77.7%	93.8%	Y	Y					Y	Y					N	Y			Y	Y	8	7	88%	N	1
Barbanti	73.6%	89.5%	Y	Y	N	N	N	N	N	Y					N	Y			Y	Y	12	6	50%	N	0
Kekata	84.9%	91.6%	Y	Y	N	Y			Y	Y	Y	Y			N	N			Y	Y	12	9	75%	N	0
Hoyt	83.7%	92.6%	Y	Y	N	Y			Y	Y	Y	Y							Y	Y	10	9	90%	N	2
Black Lake	87.2%	92.1%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	0
Lake Joseph	84.6%	93.8%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	2
Zeus	87.6%	93.3%	Y	Y	N	Y			Y	Y	Y	Y							Y	Y	10	9	90%	N	1
Ocean View	87.7%	96.5%	Y	Y			N	Y	N	Y					N	Y			Y	Y	10	7	70%	N	2
Walter Jones	90.1%	93.1%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	87.1%	94.2%	Y	Y	N				N	Y					N	N			Y	Y	9	5	56%	N	3
Chaucer	92.8%	98.5%	Y	Y	N	Y			Y	Y			Y	Y	Y	Y			Y	Y	12	11	92%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Every elementary and middle school in the sample met overall math targets.
- Seven out of eight elementary schools and all twelve middle schools with qualifying SWD subgroups failed to meet their targets for this group of students.
- All three elementary schools (Clarkson, Maryweather, and Few) and all three middle schools (McBeal, Barbanti, and Ocean View) with qualifying LEP subgroups failed to meet their targets for this group of students.
- Four elementary schools and six middle schools failed to meet their targets for their low-income subgroups.
- Seven elementary schools and nine middle schools failed to make AYP due to one or more racial/ethnic subgroups.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. The perform-

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Georgia AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	2	7
Students with limited English proficiency	3	2	3
Low-income students	15	0	4
African-American students	5	0	1
Asian/Pacific Islander students	0	0	0
Hispanic students	7	0	6
American Indian/Alaska Native students	0	0	0
White students	16	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Georgia AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	12	12	7
Students with limited English proficiency	3	3	2
Low-income students	17	6	1
African-American students	10	4	0
Asian/Pacific Islander students	1	0	0
Hispanic students	11	7	3
American Indian/Alaska Native students	0	0	0
White students	17	0	0

ance of SWDs and LEP students are proving most challenging for schools under Georgia's system. In fact, all but one elementary and more than half of the middle schools in the study with qualifying SWD subgroups failed to make AYP in reading. Every middle school in the sample with a minimum *n* of 40 in the SWD subgroup failed to make AYP in math. All but one school with a large enough LEP population to qualify as a separate subgroup failed to meet its reading targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Georgia's NCLB accountability system is, in some respects, behaving similarly as those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all make AYP in the greatest number of states—28, 22, and 21, respectively. And

Table 6. Comparisons between schools that did and didn't make AYP in Georgia, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	7	11	3	15
Average student body size	286	316	350	961
Average % low income	26	59	39	46
Average % nonwhite	29	48	51	42
Average performance†	3.51	-0.23	0.88	-0.24
Average % growth‡	113	116	110	96
Average number of targets to meet	6	9	6	11

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

these schools all made AYP in Georgia, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Georgia.

But Georgia is also home to a few anomalies. First, consider Mayberry Elementary (see Figure 3). It failed to make AYP in 19 of the 28 states in our sample, yet made AYP in Georgia. In examining Table 2, one can see that Mayberry does not meet the minimum numbers for the LEP or SWD subgroups, which creates difficulty for so many other schools within the sample. Similarly, Wolf Creek and Nemo Elementary Schools also had no SWD or LEP subgroups. With fewer accountable subgroups, and with relatively easy proficiency standards (Figure 1), these schools are able to meet AYP, even when other schools with higher average performance fail.

Second, look at Pogesto Middle School (Figure 4). Even with its relatively low average performance it made AYP in Georgia, but failed to do so in 13 of 28 states. Like Mayberry, its AYP success in Georgia is likely attributable to the relatively small number of targets (four) it has to meet (as shown in Table 3), along with the relatively easy proficiency standards in Georgia, compared to other states.

This is consistent with the patterns shown in Table 6, which compares the sample schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, elementary schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, and have fewer subgroups (meaning fewer targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would have fared under Georgia's AYP rules and (and AMOs) for 2008. We found that only 7 elementary schools and 3 middle schools—10 in all, from a sample of 36—would have made AYP in Georgia. Looking across the 28 state accountability systems examined in the study, this puts Georgia in the upper middle of the distribution in terms of the number of elementary schools making AYP (see Figure 1).

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, the NCLB accountability system in Georgia is working exactly as Congress intended: identifying as needing attention those schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Almost all the sample schools met the Georgia AMO targets for their student populations as a whole, i.e., not considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils were not meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[8] doesn't the massive failure of middle school students to meet Georgia's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the require-

[8] See footnote 4.

ments for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Idaho

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Idaho's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Idaho's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Idaho's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Idaho's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **16 of 18 elementary schools** and **all of the middle schools** in our sample failed to make AYP in 2008 under Idaho's accountability system. The high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students. It's also partly explained by Idaho's minimum subgroup size (34), which is relatively small in comparison to most other states examined in the study. This means that schools in Idaho will be accountable for more subgroups than would similar schools in other states with larger subgroup sizes.
- Looking across the 28 state accountability systems examined in the study, we find the number of elementary schools that made AYP in **Idaho was exceeded in 20 other states (Idaho ties 5 other states in having just 2 elementary schools that made AYP). Idaho joins Massachusetts, Montana, South Carolina, and North Dakota in having no middle schools that make AYP in our sample** (see Figure 1).
- Many of the schools in our sample that failed to make AYP in Idaho are meeting expected targets for their overall populations but failed because of the performance of individual subgroups, particularly students with disabilities (SWD) and English language learners.[2]
- Schools with fewer subgroups attained AYP more easily in Idaho than schools with more subgroups, even when their average student performance is much lower. In other words, **schools with greater**

Only two elementary schools and none of the middle schools in our sample made AYP in 2008 under **Idaho's** accountability system. A number of factors likely contribute to this low number. First, Idaho's minimum subgroup size is 34, which is relatively small in comparison to most other states examined in the study. This means that schools in Idaho will be accountable to more subgroups than would similar schools in other states with higher subgroup sizes. Not only did many disadvantaged subgroups fail their annual targets in Idaho, quite a few white subgroups failed as well, especially in reading. Another factor which makes it difficult for schools to make AYP in Idaho is that no confidence interval (margin of error) is applied to proficiency rate calculations.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Idaho Standards Achievement Tests.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

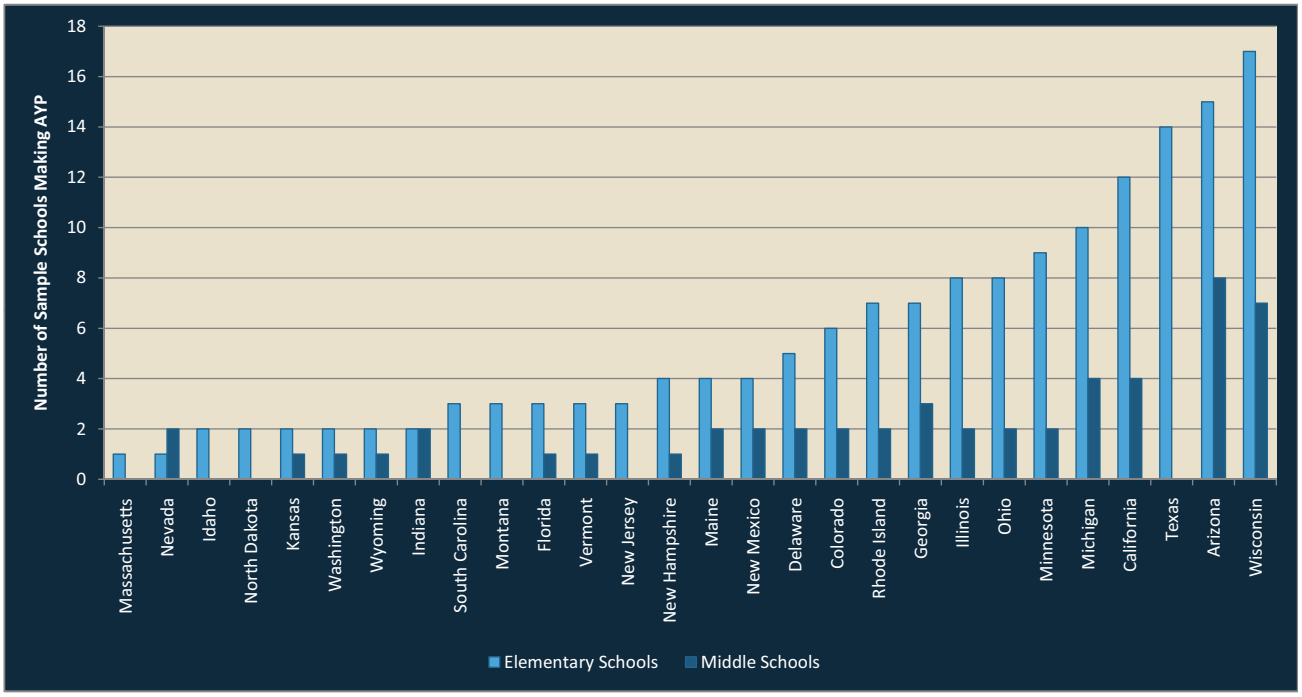

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

diversity and size face greater challenges in making AYP. This is true in other states as well.

- Middle schools have somewhat greater difficulty reaching AYP in Idaho than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is any lower than in the elementary schools.

- A strong predictor of whether or not a school will make AYP under Idaho's system is whether it has enough SWDs and enough English language learners to qualify as a separate subgroup. Every school with an SWD or limited English proficient (LEP)[3] subgroup failed to make AYP, in part because these students did not meet the state's proficiency targets in reading or math.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Idaho's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Idaho Standards Achievement Tests, the standardized state 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 tests or that allow them to take alternative assessments. In this study, however, no valid MAP scores were omitted from consideration.

plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low-income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Idaho Standards Achievement Tests (ISAT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Idaho's definitions of proficiency generally ranked about average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Idaho test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 mid-

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

[6] We gave all schools in our sample pseudonyms in this report.

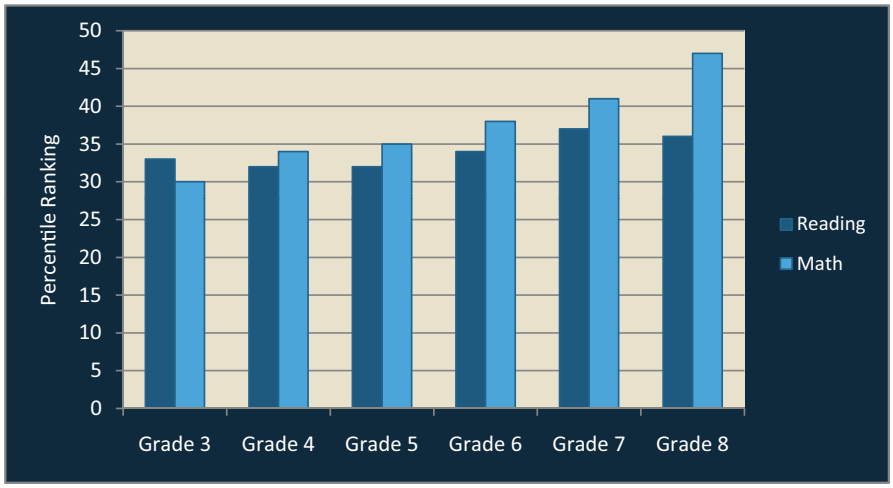

Figure 2. Idaho reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Idaho's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Idaho's cut scores are below the 50th percentile.

Table 1. Idaho AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 34	
	SWDs: 34	
	Low-income students: 34	
	LEP students: 34	
CI	**Applied to proficiency rate calculations?**	
	CI not used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	66	78
Grade 4	66	78
Grade 5	66	78
Grade 6	66	78
Grade 7	66	78
Grade 8	66	78
MATH		
Grade 3	51	70
Grade 4	51	70
Grade 5	51	70
Grade 6	51	70
Grade 7	51	70
Grade 8	51	70

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

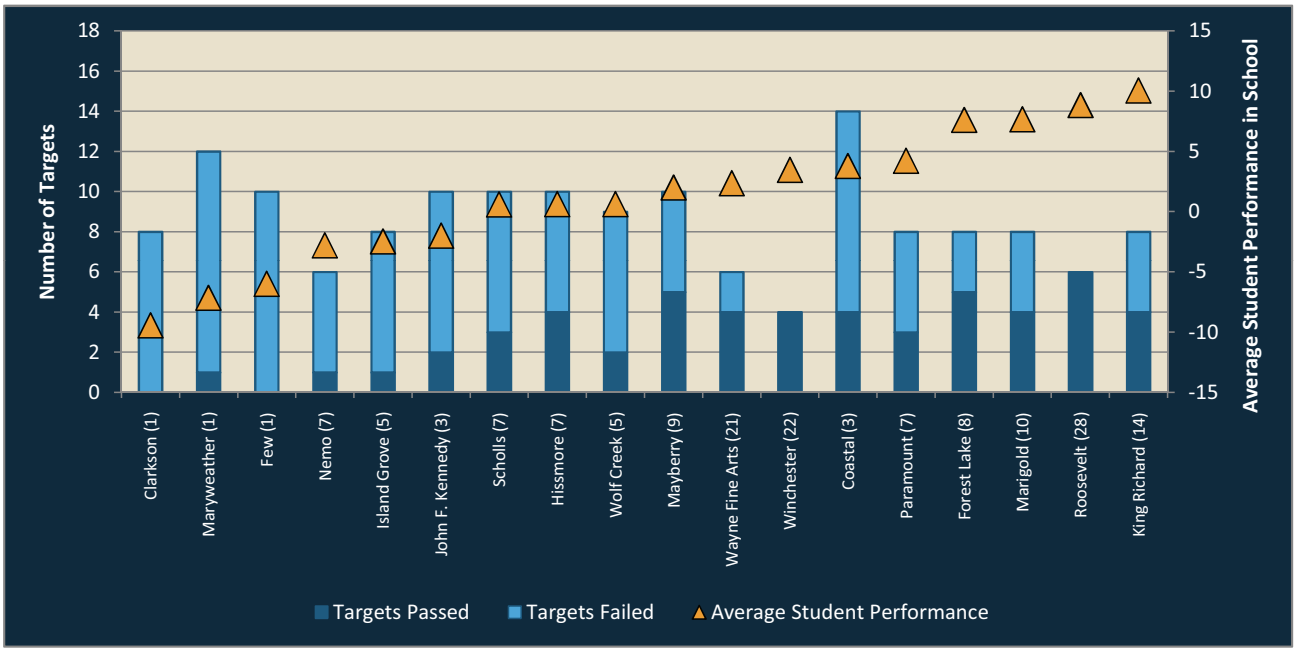

Figure 3. AYP performance of the elementary school sample under Idaho's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Idaho's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Wayne Fine Arts, for example, met four of its six targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP in the study.

dle schools would have fared under Idaho AYP rules for 2008. (In other words, the school data are from 2005–2006, as are our proficiency cut score estimates, but we are applying them against Idaho's 2008 AYP rules.)

Table 1 shows the pertinent Idaho AYP rules that we applied to elementary and middle schools in the current study. **Idaho's minimum subgroup size is 34, which is relatively small in comparison to most other states examined in the study. This means that schools in Idaho will be accountable for more subgroups than would similar schools in other states with larger subgroup sizes.**[7]

Furthermore, although the majority of states examined in the study apply confidence intervals to their student proficiency rates, Idaho does not. This means that **Idaho schools will have greater difficulty achieving their AMOs than equivalent schools in other states that report a confidence interval** around their school proficiency rates.

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their

[7] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

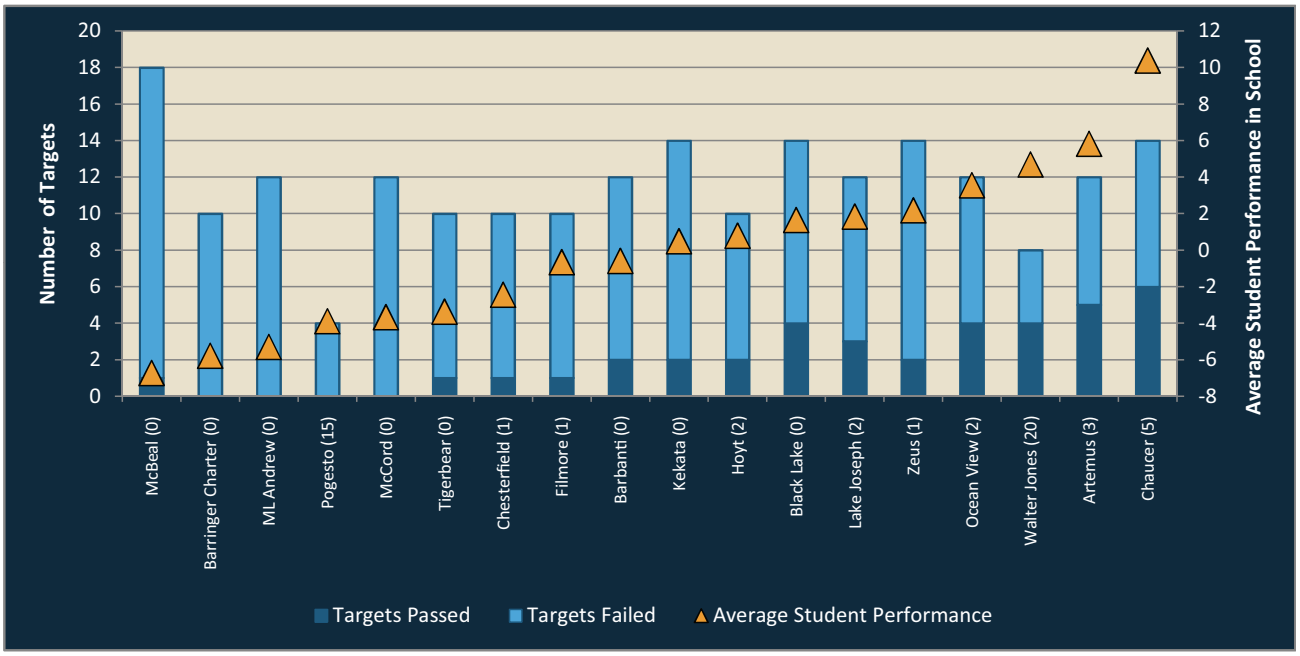

Figure 4. AYP performance of the middle school sample under Idaho's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Idaho's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that fails to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Walter Jones, for example, met four of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP

NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Idaho's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Idaho's 2008 AYP rules. **Only 2 elementary schools made AYP while 16 failed to make it.** The triangles in the figure show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The passing schools are in the right half of the figure, meaning that the highest performing students were found at these schools.

Yet, almost without regard to average student performance, the only schools that actually made AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has its own separate targets). Only Winchester and Roosevelt passed, and they had just four and six targets, respectively. Each had to make AYP for its overall student population in reading and math (two targets) and for its white population (two more targets); Roosevelt also had to make AYP for its low-income population (two targets).

Table 2. Elementary school subgroup performance of sample schools under the 2008 Idaho AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	51.1%	40.8%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	57.1%	50.7%	N	N	N	N	N	N	N	N					N	N			Y	N	12	1	8%	N	1
Few	64.6%	51.9%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Nemo	66.0%	67.0%	N	N					N	N									Y	N	6	1	17%	N	7
Island Grove	69.3%	66.7%	N	N					N	N					N	N			Y	N	8	1	13%	N	4
JFK	72.9%	60.4%	Y	N	N	N			N	N	N	N							Y	N	10	2	20%	N	3
Scholls	81.7%	68.8%	Y	N	N	N			Y	N	N	N							Y	N	10	3	30%	N	7
Hissmore	80.6%	72.1%	Y	N	N	N			Y	N	Y	N							Y	N	10	4	40%	N	7
Wolf Creek	72.5%	67.6%	Y	N		N			N	N					N	N			Y	N	9	2	22%	N	5
Alice Mayberry	77.2%	75.1%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	9
Wayne Fine Arts	79.3%	81.0%	Y	Y					N	N									Y	Y	6	4	67%	N	21
Winchester	78.8%	79.1%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	82.2%	74.8%	Y	N	N	N	N	N	Y	N	N	N			N	N			Y	Y	14	4	29%	N	3
Paramount	81.0%	76.1%	Y	N					N	N					N	N			Y	Y	8	3	38%	N	7
Forest Lake	88.5%	84.4%	Y	Y	N	N			Y	N									Y	Y	8	5	63%	N	8
Marigold	91.0%	85.6%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	10
Roosevelt	93.6%	91.5%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	89.9%	88.8%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Figure 4 illustrates the AYP performance of the sample **middle schools** under the 2008 Idaho AYP rules. **None of the 18 schools in our sample passed**—even Walter Jones, the middle school that makes AYP in the greatest number of states (20) or the school with the highest performing students (Chaucer) didn't make AYP in Idaho.

Where Do Schools Fail?

Figure 3 illustrates how some elementary schools with middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. Figures 3 and 4 do not, however, indicate which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential sub-

Table 3. Middle school subgroup performance of sample schools under the 2008 Idaho AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	48.3%	52.8%	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	18	1	6%	N	0
Barringer Charter	53.3%	57.0%	N	N	N	N			N	N	N	N			N	N					10	0	0%	N	0
ML Andrew	48.5%	56.8%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Pogesto	44.4%	61.1%	N	N															N	N	4	0	0%	N	15
McCord Charter	50.8%	60.3%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Tigerbear	60.9%	55.7%	N	N	N	N			N	N	N	N							Y	N	10	1	10%	N	0
Chesterfield	62.5%	57.8%	N	N	N	N			N	N	N	N							Y	N	10	1	10%	N	1
Filmore	62.0%	66.4%	N	N	N	N			N	N					N	N			Y	N	10	1	10%	N	1
Barbanti	59.0%	61.7%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Kekata	68.4%	66.1%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	0
Hoyt	69.2%	68.6%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	2
Black Lake	73.2%	68.8%	Y	N	N	N			N	N	N	N	Y	Y	N	N			Y	N	14	4	29%	N	0
Lake Joseph	70.1%	72.5%	Y	N	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	2
Zeus	72.2%	71.6%	Y	N	N	N	N	N	N	N	N	N			N	N			Y	N	14	2	14%	N	1
Ocean View	74.3%	81.3%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	82.0%	82.9%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	20
Artemus	81.0%	78.1%	Y	Y	N	N			N	N			Y	N	N	N			Y	Y	12	5	42%	N	3
Chaucer	83.2%	86.6%	Y	Y	N	N	N	N	N	N			Y	Y	N	N			Y	Y	14	6	43%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

groups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white.

The school-by-school findings in Tables 2 and 3 show that:

- Most elementary schools met targets in math, but not in reading, for their overall student populations.
- Almost no subgroups at the elementary or middle school level met math or reading targets, except for white youngsters in math.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. We see that every school with large enough populations of students with disabilities, LEP, Hispanic, or American Indian/Alaska Natives to qualify as separate subgroups failed to meet its reading and math targets for these students. In fact, the only subgroups where *any* schools in the sample met their targets in both reading and math were Asian/Pacific Islander, and white.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 APY Idaho rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	11	10	11
Students with limited English proficiency	4	4	4
Low-income students	17	11	16
African-American students	5	3	5
Asian/Pacific Islander students	0	0	0
Hispanic students	7	7	7
American Indian/Alaska Native students	0	0	0
White students	16	0	7

Table 5. Summary of subgroup performance of sample middle schools under 2008 APY Idaho rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	7	7	7
Low-income students	17	17	17
African-American students	10	10	10
Asian/Pacific Islander students	4	1	2
Hispanic students	14	14	14
American Indian/Alaska Native students	1	1	1
White students	17	3	9

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Idaho's NCLB accountability system (at least in terms of the elements studied here) failed more schools than do most states in our sample. In fact, only two states (Massachusetts and Nevada) failed more elementary schools than Idaho. Similarly, Idaho is one of only five states (along with Massachusetts, South Dakota, Montana, and North Dakota) that have zero passing middle schools in our sample (see Figure 1).

The most likely explanation for the difference in Idaho, relative to the other states examined, has to do with Idaho's more stringent AYP rules. Defining subgroups at 34 means that a school in Idaho will have more subgroups and consequently more chances to fail to make

Table 6. Comparisons between schools that did and didn't make AYP in Idaho, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	2	16	0	18
Average student body size	225	315	n/a	859
Average % low income	13	50	n/a	45
Average % nonwhite	25	43	n/a	44
Average performance†	6.16	0.61	n/a	-0.05
Average % growth‡	121	114	n/a	98
Average number of targets to meet	5	9	n/a	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

AYP.[8] Additionally, Idaho does not use a confidence interval as do most of the other states examined. This means that its schools have a more difficult time meeting their targets compared to states that use confidence intervals.

This is consistent with the patterns shown in Table 6, which compares schools making and not making AYP on several academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have smaller student populations, fewer subgroups (and thus fewer targets to meet), and lower percentages of low-income students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Idaho's AYP rules (and AMOs) for 2008. We found that only 2 elementary schools and no middle schools—2 in all from a sample of 36—would have made AYP in Idaho. Looking across the 28 state accountability systems, this puts Idaho near the lower end of the sample distribution in terms of the number of schools making AYP (see Figure 1).

Several other factors are important to note for Idaho. First, Idaho's minimum subgroup size is relatively small in comparison to most other states examined in the study, meaning that schools in Idaho will be accountable for more subgroups than would similar schools in other states with higher subgroup sizes. Finally, Idaho, unlike most other states, does not apply confidence intervals so schools will have greater difficulty achieving their annual targets than equivalent schools in other states that report a confidence interval.

The overriding goal of the federal NCLB is to eliminate educational disparities within and across states; it is important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Idaho's NCLB accountability system is working exactly as Congress intended:

[8] It is worth noting, however, that schools in Idaho are likely to be small and an *n* size of 34 probably makes sense.

identifying as needing attention those schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. A moderate number of schools made annual targets in Idaho for their student populations as a whole, that is, without considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils were not meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[9] doesn't the failure of these students to meet Idaho's targets (especially at the middle school level) indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[9] See footnote 4.

Illinois

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines the NCLB accountability system in Illinois—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough the Illinois system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under the Illinois system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against the Illinois AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **10 out of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under the Illinois accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, **we find that only six states exceed Illinois in terms of the number of elementary schools making AYP (Minnesota, Michigan, California, Texas, Arizona, and Wisconsin)(see Figure 1)**. Illinois ties with Ohio, each with eight (out of 18) elementary schools making AYP.
- Nearly all the schools in our sample that failed to make AYP in Illinois are meeting expected targets for their overall populations but failing because of the performance of individual subgroups.[2]
- Two sample schools made AYP in Illinois that failed to make AYP in most other states. This is likely because the **proficiency standards in Illinois are relatively easy compared to those in other states**; these schools also have fewer accountable subgroups (**the minimum *n* size in Illinois is a bit higher than in many other states, meaning it takes more students within a subgroup category for that group to be held separately accountable**).

There are several contributing factors to the relatively high number of sample schools that make AYP in **Illinois**. The first is the state's low proficiency standards (or cut scores). In reading, cut scores range from the 35th to the 22nd percentile, and in math, cut scores range from the 20th to the 15th percentile. Moreover, schools have fewer accountable subgroups in Illinois than in most other states because Illinois's minimum subgroup size (45) is a bit higher than the minimum in many other states. In other words, Illinois schools must have more students within a subgroup category in order for that group to be held separately accountable. With fewer accountable subgroups and relatively easy proficiency standards, it's possible for more study schools to make AYP in Illinois.

[1] A cut score is the minimum score a student must receive on NEWA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Illinois Standards Achievement Test.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

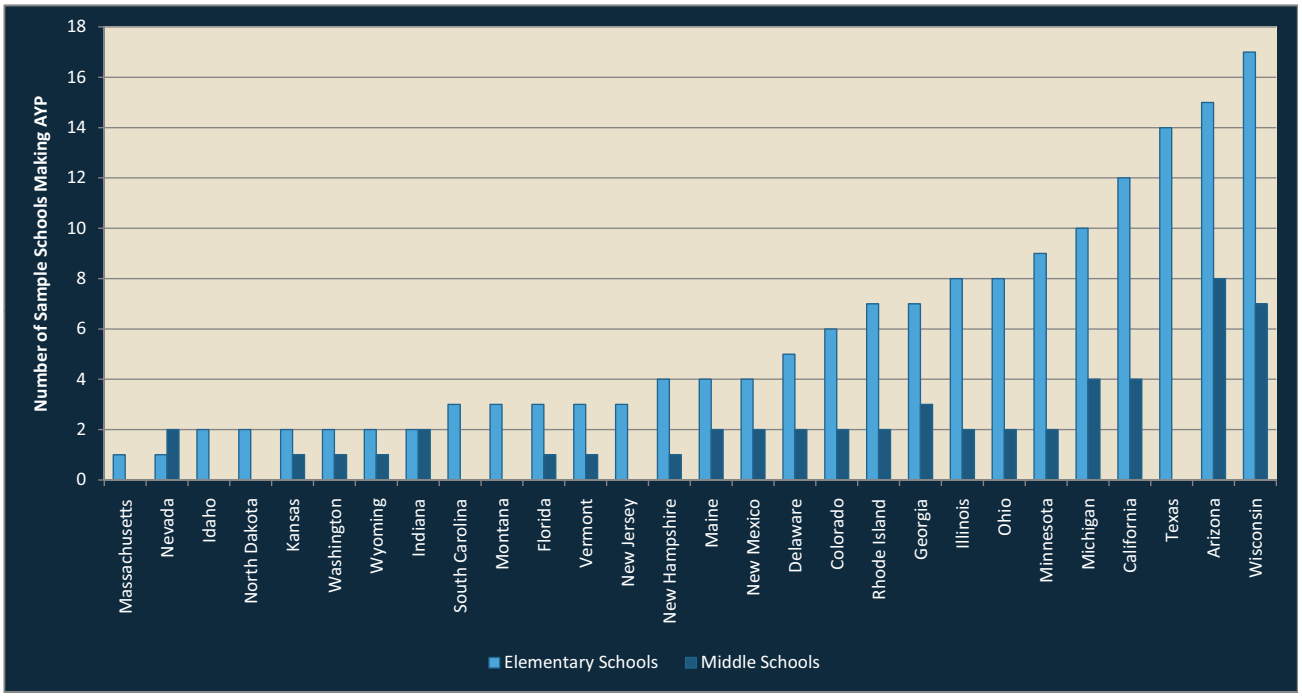

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- As in other states, schools with fewer subgroups attain AYP more easily in Illinois than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.
- Middle schools have greater difficulty reaching AYP in Illinois than do elementary schools, primarily because their student populations are larger and they therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.
- A strong predictor of whether or not a school will make AYP under the Illinois system is whether it has enough students with disabilities (SWD) or English language learners to qualify as a separate subgroup. All but one school with limited English proficient (LEP)[3] or SWD subgroups failed to make AYP, in part because these students did not meet the state's annual proficiency targets, especially in reading.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Illinois' tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Illinois Standards Achievement Test, the standardized state test. 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.

profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Illinois Standards Achievement Test (ISAT) were taken from *The Proficiency Illusion* (as shown in Figure 2), which found that the definitions of reading and math proficiency in Illi-

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

[6] We gave all schools in our sample pseudonyms in this report.

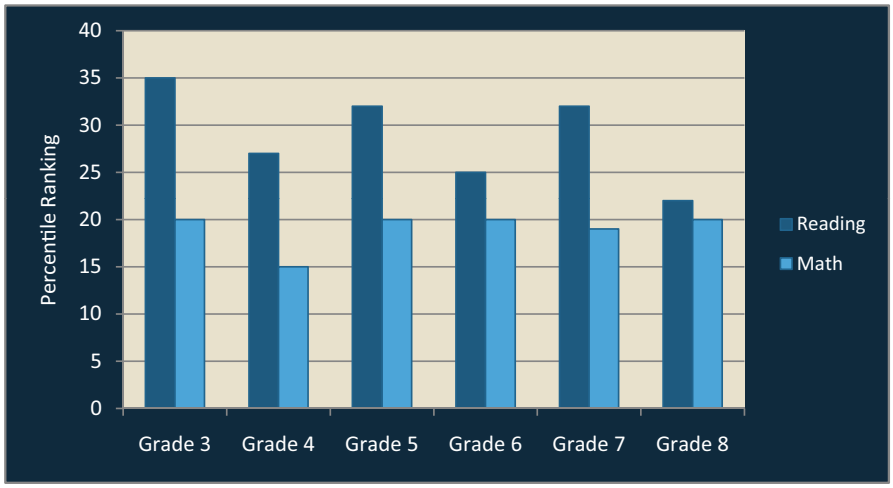

Figure 2. Illinois reading and math cut score estimates, expressed as percentile ranks (2006))

Note: This figure illustrates the difficulty of the cut scores (or proficiency passing scores) in Illinois for the state's reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of the state's cut scores are at or below 35th percentile.

nois were generally below average, or among the less difficult of the 26 states examined in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Illinois test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Illinois AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from 2005–2006, but we are applying them against the Illinois 2008 AYP rules.

Table 1 shows the pertinent Illinois AYP rules that were applied to elementary and middle schools in the current study. The state's minimum subgroup size is 45, which is slightly larger than most other states we examined. Most states examined also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. So, although schools are supposed to get 62.5% of their students to the proficient level on the state reading test, as well as 62.5% of their students in each subgroup, applying the confidence interval means that the real targets can actually be lower (particularly with smaller groups). In other words, using a confidence interval makes it easier for Illinois schools to meet the targets as defined by their AMOs.[7]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population

[7] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

Table 1. Illinois AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 45	
	SWDs: 45	
	Low-income students: 45	
	LEP students: 45	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	47.5	62.5
Grade 4	47.5	62.5
Grade 5	47.5	62.5
Grade 6	47.5	62.5
Grade 7	47.5	62.5
Grade 8	47.5	62.5
MATH		
Grade 3	47.5	62.5
Grade 4	47.5	62.5
Grade 5	47.5	62.5
Grade 6	47.5	62.5
Grade 7	47.5	62.5
Grade 8	47.5	62.5

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under the Illinois' AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under the 2008 AYP rules in Illinois. **Eight elementary schools made AYP and ten failed to make it**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student and positive values indicating above-grade-level performance. Six of the eight passing schools are on the right half of the figure, meaning that students with the highest average academic performance were found at these schools.

Yet almost without regard to average student performance, the schools that made AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (since each subgroup has its own separate targets). For example, Wayne Fine Arts and Winchester passed, but had only four targets each. Each school must make AYP for its overall student population in reading and math (two targets) and for its white population, resulting in four total targets.

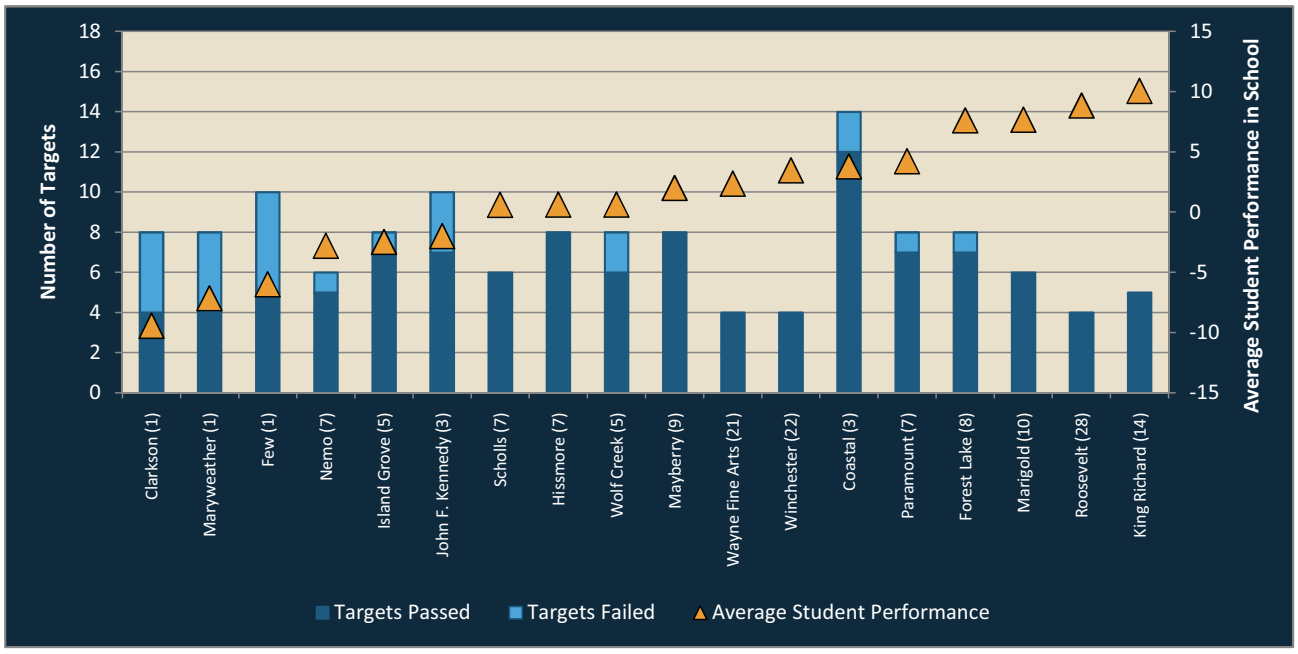

Figure 3. AYP performance of the elementary school sample under the Illinois 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under the Illinois AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Wolf Creek Elementary, for example, met six of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Illinois AYP rules. **Of 18 middle schools in our sample, only 2 passed**—1 low-performance school (Pogesto) and 1 high-performance school (Walter Jones), both of which have relatively few qualifying subgroups and consequently few targets to meet.

Figure 5 indicates the degree to which elementary schools' overall math proficiency rates are aided by the confidence interval. On this figure, the darker portions of the bars show the actual proficiency rates at each school and the lighter portions of the bars show the degree to which the proficiency rates were increased by applying the confidence interval. The orange lines show the AMOs needed to meet AYP. This figure shows that none of the sample elementary schools was assisted by the confidence interval, because **the math targets in Illinois are low relative to the schools' overall performance**. The pattern is much the same for middle school math proficiency rates and for reading proficiency rates at the elementary and middle school levels (not shown). In fact, only one elementary school, John F. Kennedy, received assistance from the confidence interval in meeting its reading target. As shown in Figure 3, however, this school still failed to meet all its subgroup targets. **Overall, applying the confidence interval had little or no effect on whether schools made AYP in Illinois**.[8]

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet, thanks to fewer subgroups. These figures do not, however, indicate which subgroups

[8] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figure 5. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

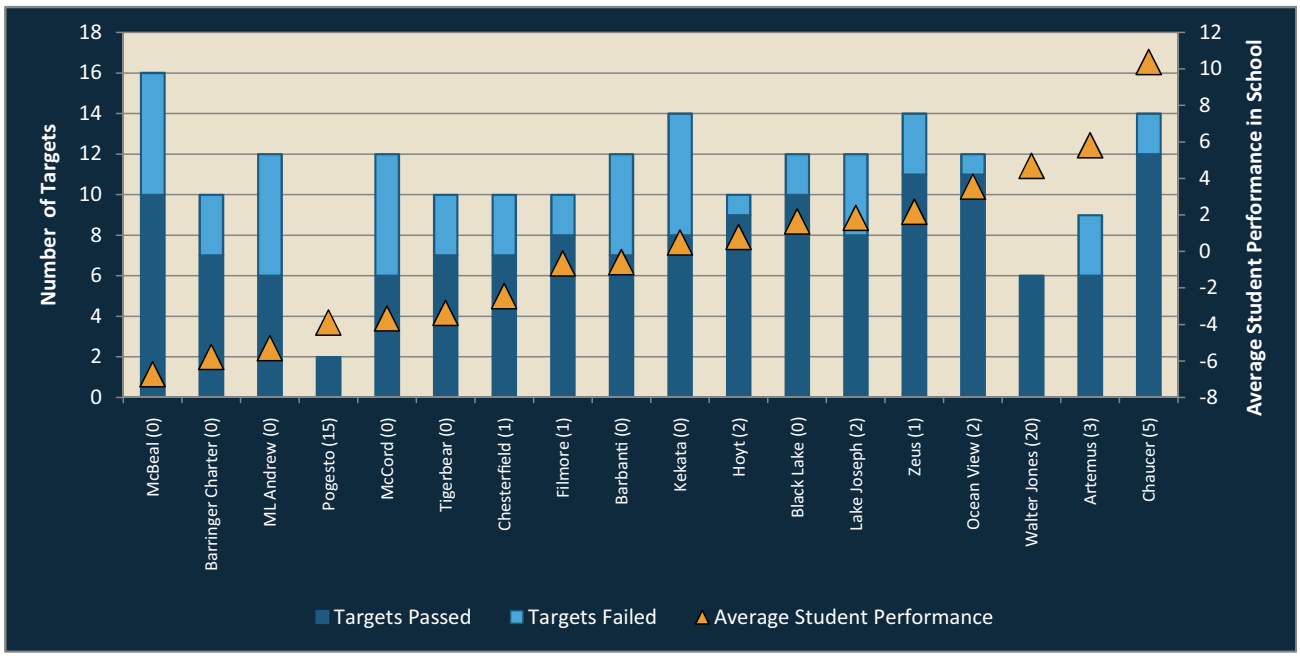

Figure 4. AYP performance of the middle school sample under the Illinois 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under the AYP rules in Illinois (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them(light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Artemus Middle School, for example, met six of its nine targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

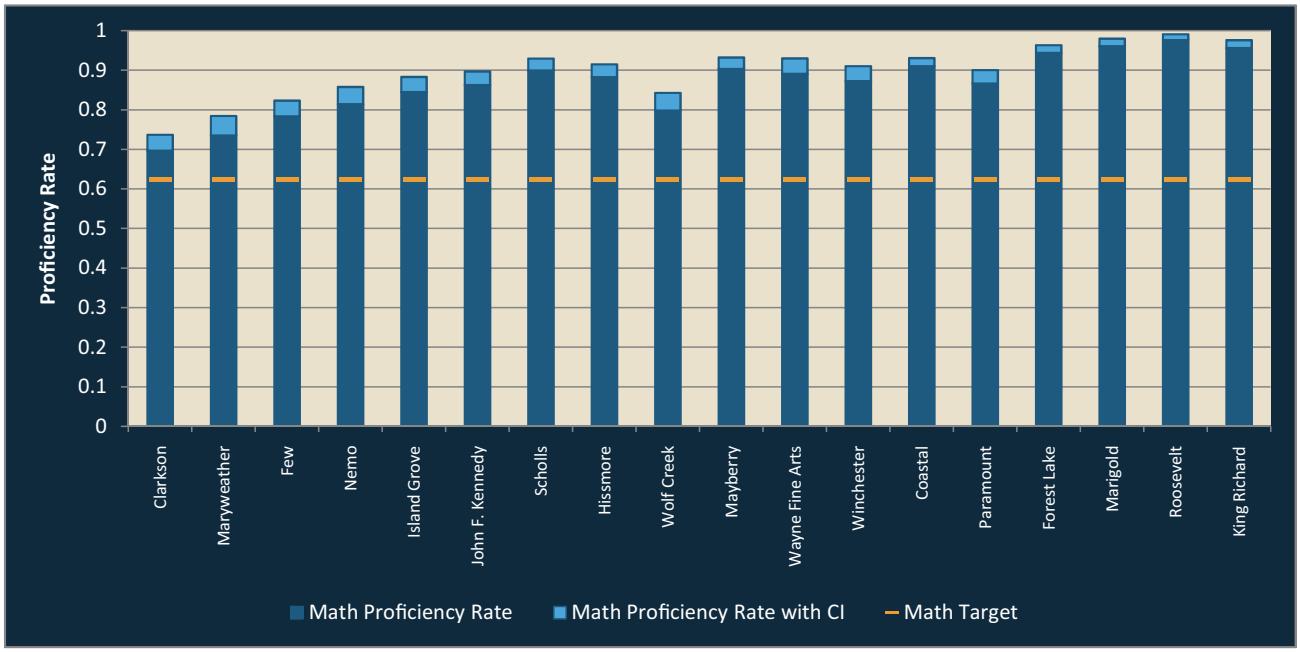

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Illinois AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	69.7%	40.8%	Y	N			Y	N	Y	N					Y	N					8	4	50%	N	1
Maryweather	73.5%	52.1%	Y	N			Y	N	Y	N					Y	N					8	4	50%	N	1
Few	78.4%	51.9%	Y	N	Y	N	Y	N	Y	N					Y	N					10	5	50%	N	1
Nemo	81.4%	68.4%	Y	Y					Y	N									Y	Y	6	5	83%	N	7
Island Grove	84.5%	68.3%	Y	Y					Y	Y					Y	N			Y	Y	8	7	88%	N	4
JFK	86.2%	61.6%	Y	Y	Y	N			Y	N	Y	N							Y	Y	10	7	70%	N	3
Scholls	89.9%	69.9%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Hissmore	88.2%	73.3%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Wolf Creek	79.8%	68.9%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	5
Alice Mayberry	90.3%	74.0%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	89.1%	81.0%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	87.3%	79.6%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	91.0%	76.5%	Y	Y	Y	N	Y	N	Y	Y	Y	Y			Y	Y			Y	Y	14	12	86%	N	3
Paramount	86.6%	76.5%	Y	Y					Y	N					Y	Y			Y	Y	8	7	88%	N	7
Forest Lake	94.4%	84.9%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	96.0%	86.3%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	10
Roosevelt	97.6%	92.5%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	95.6%	89.1%	Y	Y	Y														Y	Y	5	5	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Illinois rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Three elementary schools (Clarkson, Maryweather, and Few) failed to meet reading targets for their overall school population.

Table 3. Middle school subgroup performance of sample schools under the 2008 Illinois AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	69.9%	62.6%	Y	Y	N	N	N	N	Y	N	Y	Y			Y	N	Y	Y	Y	Y	16	10	63%	N	0
Barringer Charter	76.2%	63.2%	Y	Y	N	N			Y	Y	Y	N			Y	Y					10	7	70%	N	0
ML Andrew	73.3%	66.9%	Y	Y	N	N			N	N	N	N			Y	Y			Y	Y	12	6	50%	N	0
Pogesto	74.1%	74.1%	Y	Y																	2	2	100%	Y	15
McCord Charter	74.7%	69.6%	Y	Y	N	N			N	N	N	N			Y	Y			Y	Y	12	6	50%	N	0
Tigerbear	79.1%	66.8%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	0
Chesterfield	83.6%	70.7%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	1
Filmore	82.9%	76.8%	Y	Y	N	N			Y	Y					Y	Y			Y	Y	10	8	80%	N	1
Barbanti	77.8%	71.7%	Y	Y	N	N	N	N	Y	N					Y	Y			Y	Y	12	7	58%	N	0
Kekata	86.0%	73.9%	Y	Y	N	N	N	N	Y	Y	Y	N			Y	N			Y	Y	14	8	57%	N	0
Hoyt	88.4%	77.6%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	2
Black Lake	89.2%	78.6%	Y	Y	N	N			Y	Y	Y	Y			Y	Y			Y	Y	12	10	83%	N	0
Lake Joseph	86.6%	81.3%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	89.9%	79.3%	Y	Y	Y	N	Y	N	Y	Y	Y	Y			Y	N			Y	Y	14	11	79%	N	1
Ocean View	90.6%	87.5%	Y	Y	Y	Y	Y	N	Y	Y					Y	Y			Y	Y	12	11	92%	N	2
Walter Jones	91.3%	85.1%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	92.0%	84.4%	Y	Y		N			Y	N					Y	N			Y	Y	9	6	67%	N	3
Chaucer	94.3%	90.8%	Y	Y	Y	N	Y	N	Y	Y			Y	Y	Y	Y			Y	Y	14	12	86%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

- No elementary schools failed to meet math targets.
- All sample middle schools met overall targets in both reading and math.
- One elementary school (Forest Lake) and three middle schools (Filmore, Hoyt, and Black Lake) missed only because of the SWD subgroup.
- One middle school (Ocean View) failed only because of its LEP subgroup.
- Two elementary schools (Nemo and Paramount) failed only because of their low-income subgroup.
- One elementary school (Island Grove) passed in every subgroup except for Hispanic students.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, elementary students did well on the state math test, perhaps because the Illinois proficiency cut scores are easier in math than in reading at the elementary grades (see Figure 2). Second, the performance of SWDs and LEP students are proving most challenging for schools under the Illinois system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Illinois AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	5	0	4
Students with limited English proficiency	4	0	4
Low-income students	14	0	7
African-American students	4	0	1
Asian/Pacific Islander students	0	0	0
Hispanic students	7	0	5
American Indian/Alaska Native students	0	0	0
White students	15	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Illinois AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	15	11	15
Students with limited English proficiency	7	4	7
Low-income students	17	2	5
African-American students	10	2	6
Asian/Pacific Islander students	1	0	0
Hispanic students	13	0	4
American Indian/Alaska Native students	1	0	0
White students	16	0	0

minimum *n* of 45. Students with LEP are also struggling to meet the state's targets; every single school with a large enough LEP population to qualify as a separate subgroup fails to meet its reading targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that the NCLB accountability system in Illinois is, in many respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Illinois, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Illinois.

Table 6. Comparisons between schools that did and didn't make AYP in Illinois, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	8	10	2	16
Average student body size	255	344	124	951
Average % low income	34	56	42	45
Average % nonwhite	32	48	27	46
Average performance†	4.45	-1.36	0.40	-0.11
Average % growth‡	113	117	109	97
Average number of targets to meet	6	9	4	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

But Illinois is also home to a few anomalies. First, consider Mayberry Elementary (see Figure 3). It failed to make AYP in 19 of the 28 states in our sample, yet made AYP in Illinois. In examining Table 2, one can see that Mayberry didn't meet the state's minimum numbers for the LEP or SWD subgroups, which created difficulty for so many other schools within the sample. This is likely a reflection of the fact that the Illinois minimum subgroup size, 45, is a bit higher than in many other states. **In other words, Illinois schools must have more students within a subgroup category for that group to be held separately accountable. With fewer accountable subgroups, and with relatively easy proficiency standards (Figure 2), Mayberry made AYP, even when other schools with higher average performance failed.**

Second, look at Pogesto Middle School (Figure 4). Even with its relatively low average performance, it makes AYP in Illinois, but fails to do so in 13 of 28 states. Like Mayberry, its AYP success in Illinois is likely attributable to the relatively small number of targets (two) it has to meet, as shown in Table 3, along with the relatively easy proficiency standards in Illinois, compared to other states.

This is consistent with the patterns shown in Table 6, which compares the schools that did and didn't make AYP on several academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and lower percentages of low-income students. Similarly, middle schools that made AYP have slightly higher performing students, on average, than middle schools that didn't, but have dramatically smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would have fared under the Illinois AYP rules (and AMOs) for 2008. We found that 8 elementary schools and 2 middle schools—10 in all from a sample of 36—would have made AYP in Illinois. Looking across the 28 state accountability systems examined in the study, this puts Illinois towards

the upper end of the distribution in terms of the number of schools making AYP (as shown in Figure 1).

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, the NCLB accountability system in Illinois is working exactly as Congress intended: identifying as needing attention those schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Almost all the sample schools met the Illinois AMO targets for their student populations as a whole, i.e., not considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils were not meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[9] does the massive failure of middle school students to meet Illinois targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

[9] See footnote 4.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Indiana

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Indiana's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Indiana's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Indiana's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Indiana's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **16 of 18 elementary schools** and **16 of 18 middle schools in our sample failed to make AYP** in 2008 under Indiana's accountability system. (This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-income students.[2])
- Looking across the 28 state accountability systems examined in the study (see Figure 1), only two states passed fewer of the sample elementary schools than Indiana. Indiana **tied with 5 other states with just 2 elementary schools making AYP**.
- Many of the schools in our sample that failed to make AYP in Indiana met expected targets for their overall populations but didn't make AYP because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[3]
- In Indiana, schools with fewer subgroups attained AYP more easily than schools with more subgroups, even when their average student performance was much lower. In other words, schools with greater diversity and size face greater challenges in making AYP. This is the case in other states as well.
- A strong predictor of whether or not a school will make AYP under Indiana's system is whether it has enough SWDs or English language learners to qualify

Indiana has fewer schools making AYP than in many other states in our study. This is particularly interesting because Indiana's definitions of proficiency in reading and math generally ranked below the average compared with the standards set by the other states. However, Indiana's annual targets in reading (the percentage of students in various subgroups that have to meet proficiency) are relatively difficult to achieve. Specifically, 72.4 percent of a given population in any school would have to be proficient on the state reading exam for the school to make AYP in 2008. In addition, Indiana's minimum subgroup size is smaller than most other states', meaning that schools in Indiana are accountable for more subgroups than similar schools in other states.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Indiana Statewide Testing for Educational Progress Plus.

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

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

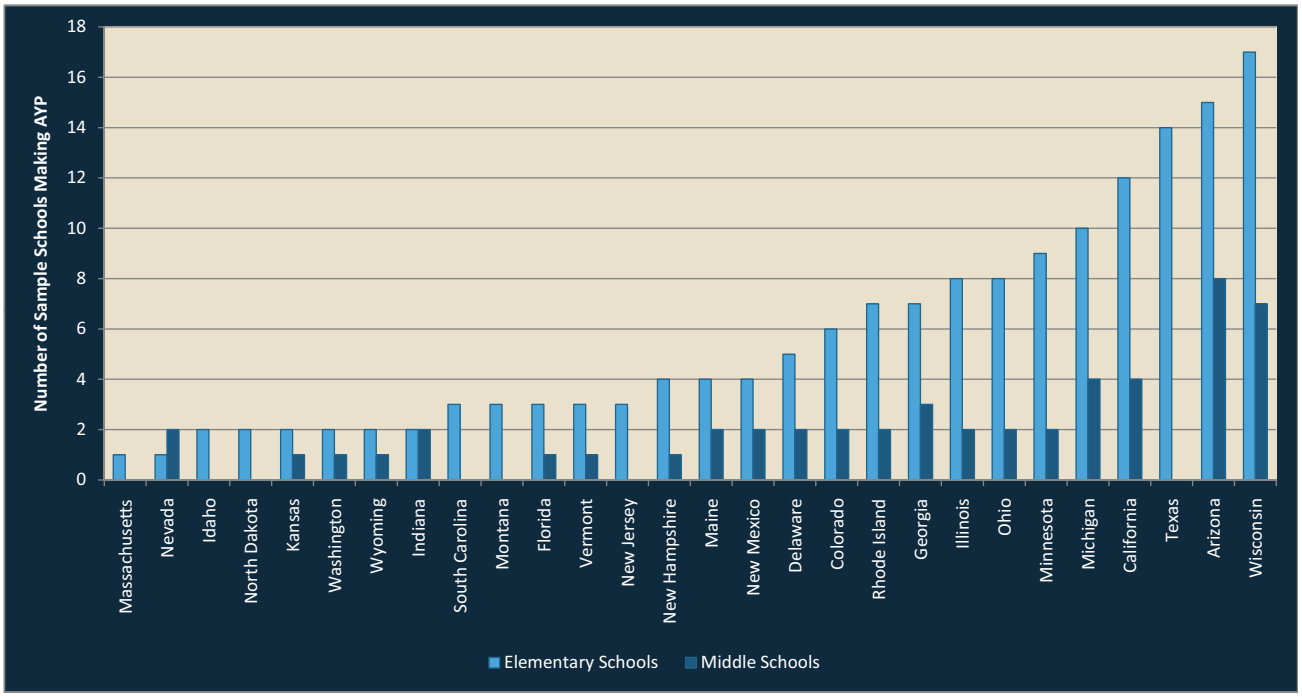

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

as a separate subgroup. Every school with limited English proficient (LEP)[4] subgroups and SWDs failed to make AYP, in part because these students did not meet the state's targets in reading and/or math.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Indiana's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools

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

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Indiana Statewide Testing for Educational Progress-Plus, the standardized state test. 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.

with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Indiana's definitions of proficiency generally ranked slightly below the average compared with the standards set by the other 25 states in that study. These cut score were used to estimate whether students would have scored as proficient or better on the Indiana test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Indiana AYP rules for 2008. (In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Indiana's 2008 AYP rules.)

Table 1 shows the pertinent Indiana AYP rules that were applied to elementary and middle schools in this study. **Indiana's minimum subgroup size is 30, which is smaller than most other states we examined.**[7]

Although most states examined also apply confidence intervals (or margins of statistical error) to their measure-

[6] We gave all schools in our sample pseudonyms in this report.

[7] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

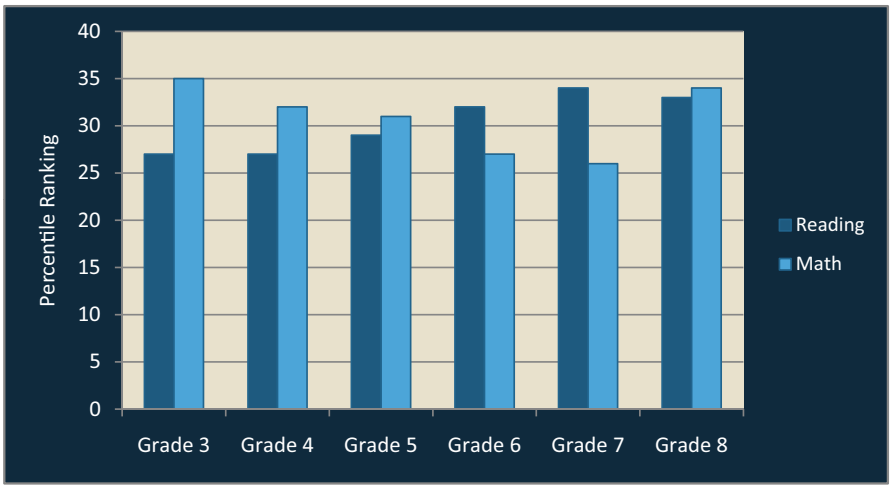

Figure 2. Indiana reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Indiana's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Indiana's cut scores are at or below the 35th percentile.

Table 1. Indiana AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 30	
	Low-income students: 30	
	LEP students: 30	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	58.8	72.4
Grade 4	58.8	72.4
Grade 5	58.8	72.4
Grade 6	58.8	72.4
Grade 7	58.8	72.4
Grade 8	58.8	72.4
MATH		
Grade 3	57.1	71.4
Grade 4	57.1	71.4
Grade 5	57.1	71.4
Grade 6	57.1	71.4
Grade 7	57.1	71.4
Grade 8	57.1	71.4

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

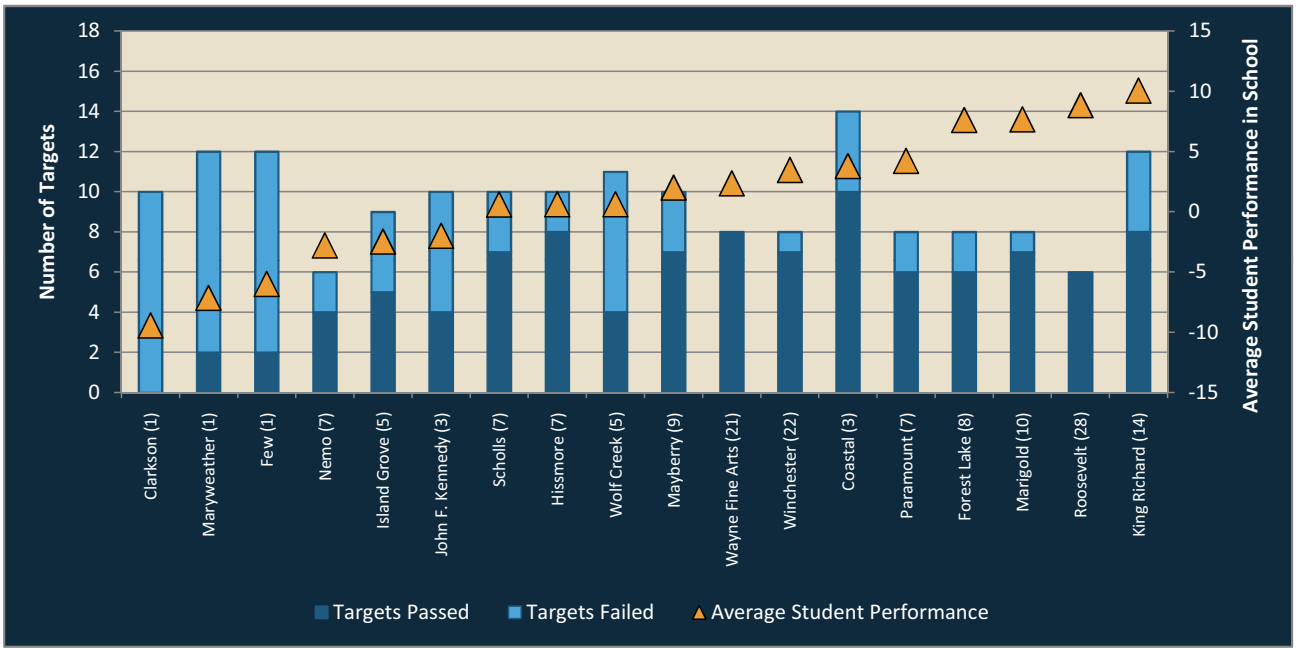

Figure 3. AYP performance of the elementary school sample under the Indiana 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Indiana's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not make them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed to make AYP. Marigold Elementary, for example, met seven of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

ments of student proficiency rates, **Indiana's 99% confidence interval gives schools greater leniency than the 95% confidence interval used by most other states**. So, for instance, although schools are supposed to get 72.4% of their students to the proficient level on the state reading test (and 72.4% of their students in each subgroup), applying the confidence interval means that the real target can actually be lower, particularly with smaller groups.[8]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population

[8] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

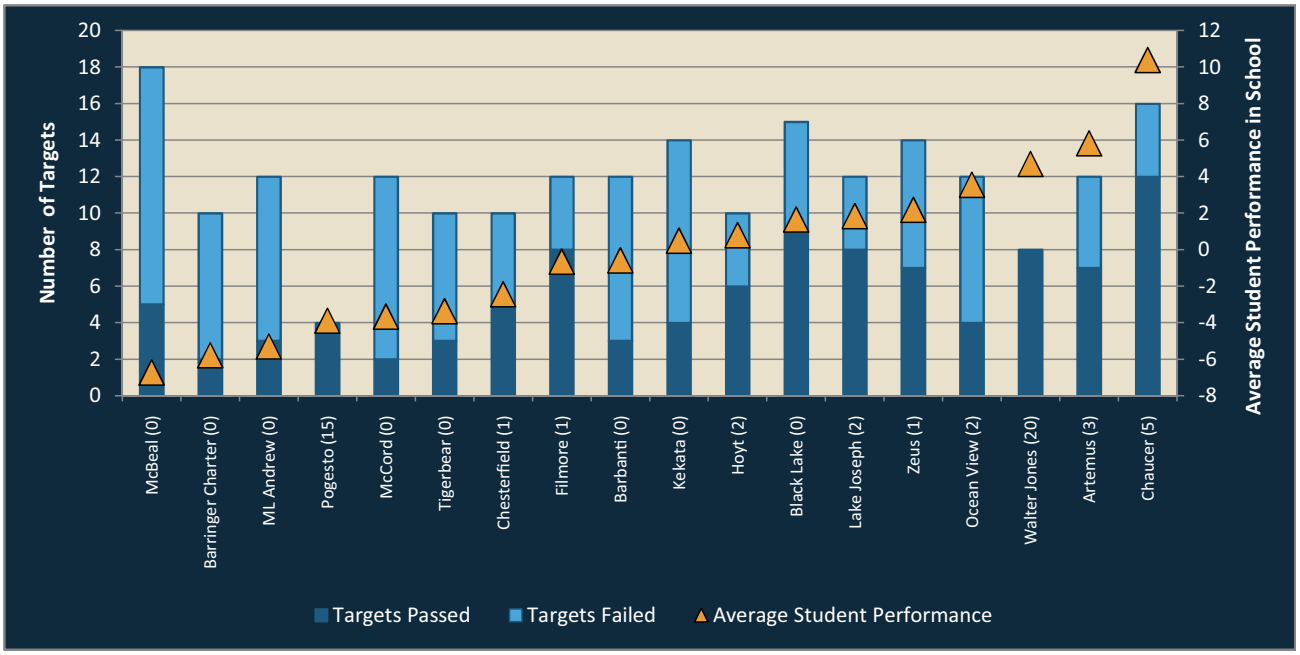

Figure 4. AYP performance of the middle school sample under the Indiana 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Indiana's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed to make AYP. Chaucer Middle School, for example, meets 12 of its 16 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

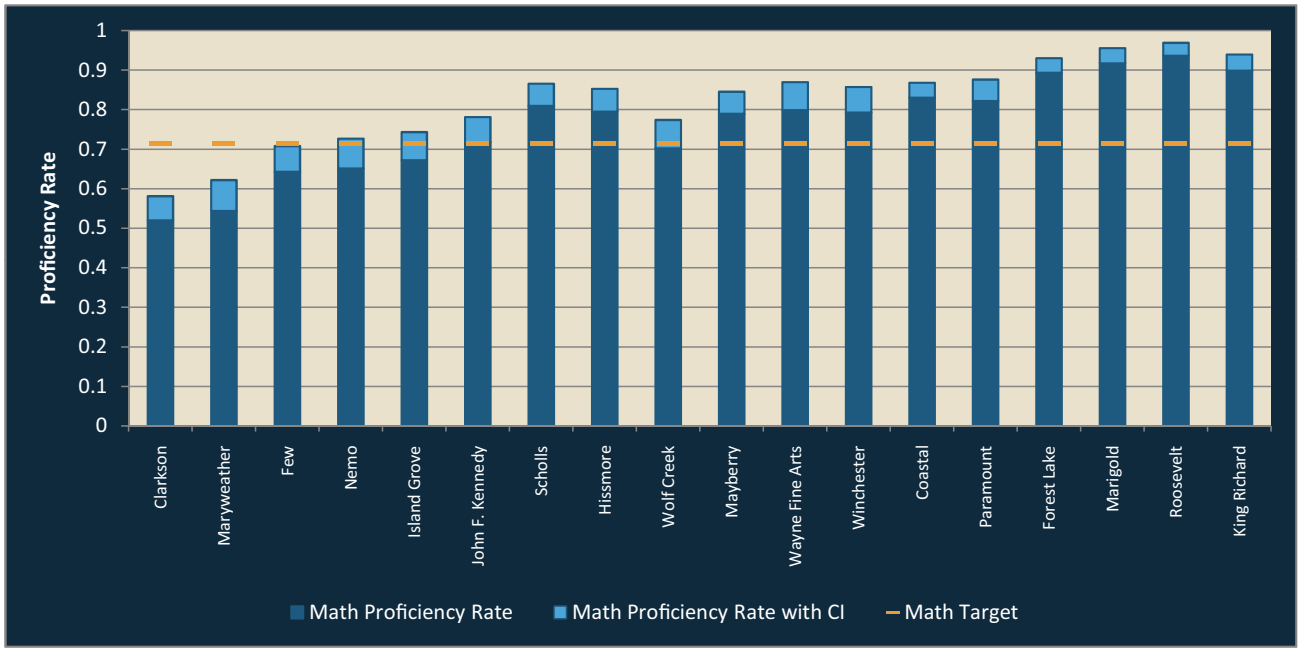

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample elementary schools (Nemo and Island Grove) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

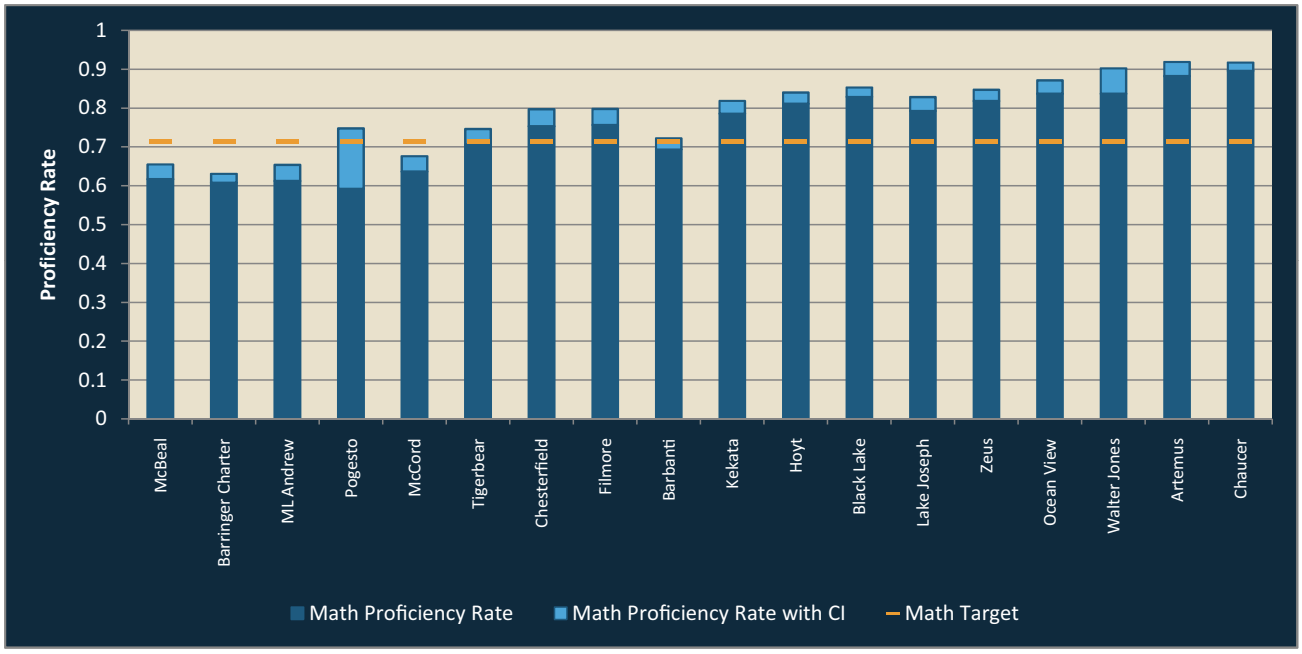

Figure 6. Impact of the confidence interval on middle school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools (Pogesto) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion. .

and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Indiana's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Indiana's 2008 AYP rules. **Only 2 elementary schools out of 18 made AYP**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The two schools that made AYP are in the right half of the figure, meaning that the highest average performing students were found at these schools.

But among the schools in the right half of the figure, the ones that made AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet (since each subgroup has its own separate targets). For example, Wayne Fine Arts and Roosevelt made AYP, but had only eight and six targets each, respectively. Each school must make AYP for its overall student population in reading and math (two targets), for its low-income students (two targets), and for its white population (two more targets). Wayne Fine Arts also has to make AYP for its African American population (two targets).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Indiana AYP rules. **Of 18 middle schools in our sample, only 2 made AYP** – one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the annual measurable objective (or annual

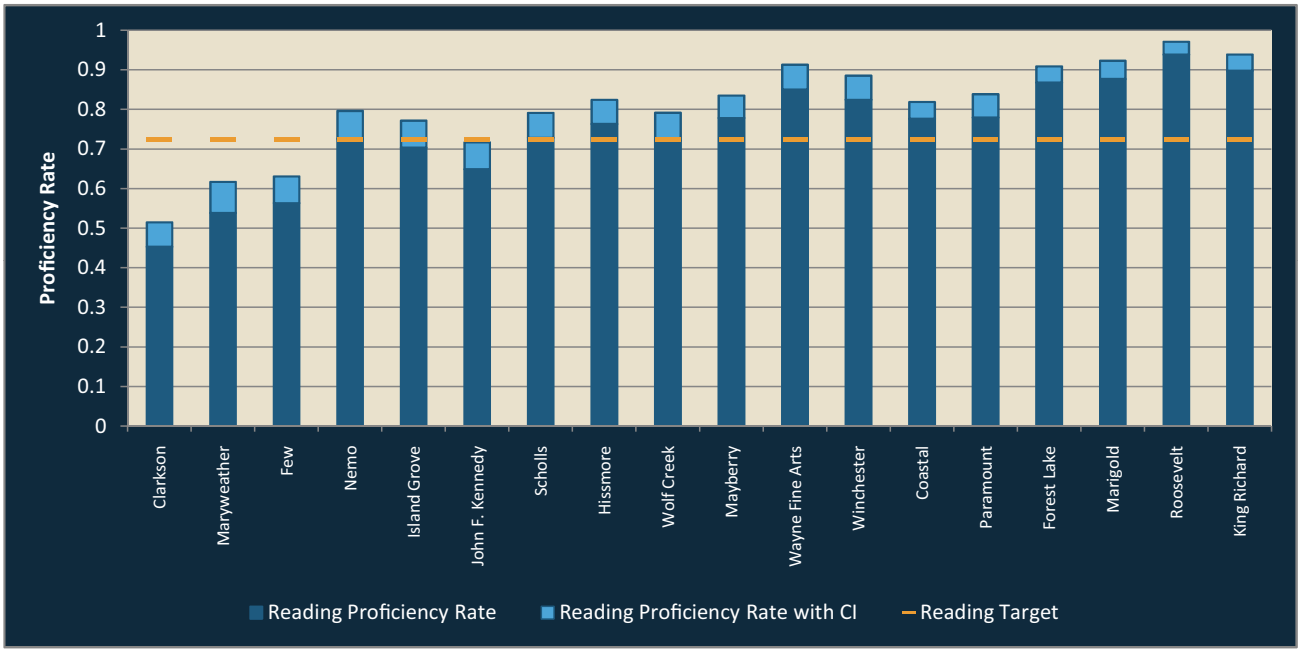

Figure 7. Impact of the confidence interval on elementary school reading proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (Island Grove) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

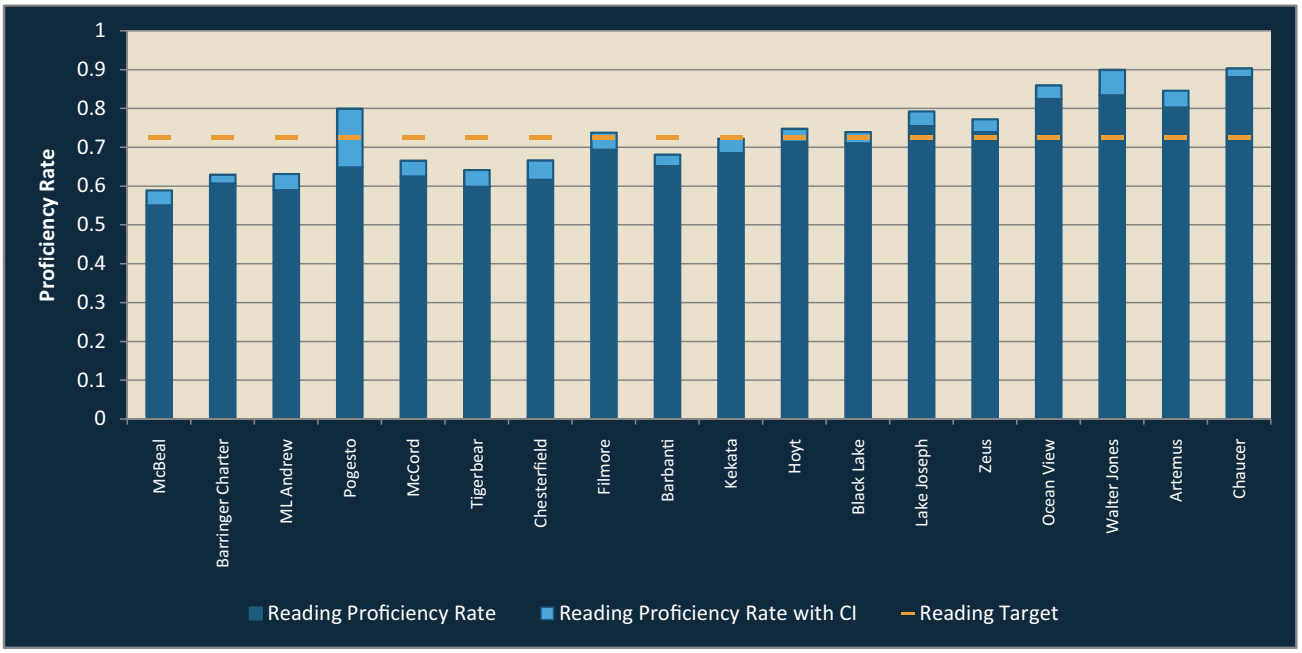

Figure 8. Impact of the confidence interval on middle school reading proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample middle schools (Pogesto and Filmore) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Indiana AYP

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	52.0%	45.4%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Maryweather	54.3%	53.9%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	1
Few	64.3%	56.4%	N	N	N	N	N	N	N	N					Y	N			Y	N	12	2	17%	N	1
Nemo	65.1%	72.6%	Y	Y					N	N									Y	Y	6	4	67%	N	7
Island Grove	67.2%	70.4%	Y	Y				N	N	Y					N	N			Y	Y	9	5	56%	N	4
JFK	71.7%	64.9%	Y	N	N	N			Y	N	N	N							Y	Y	10	4	40%	N	3
Scholls	81.0%	72.9%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	7
Hissmore	79.5%	76.3%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	70.2%	72.1%	Y	Y	N	N		N	N	N					N	N			Y	Y	11	4	36%	N	5
Alice Mayberry	79.0%	77.9%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	9
Wayne Fine Arts	79.9%	85.1%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	79.2%	82.5%	Y	Y	Y	N									Y	Y			Y	Y	8	7	88%	N	22
Coastal	83.0%	77.7%	Y	Y	N	N	Y	N	Y	Y	Y	N			Y	Y			Y	Y	14	10	71%	N	3
Paramount	82.2%	78.0%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	89.3%	86.8%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	91.7%	87.7%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	10
Roosevelt	93.6%	93.9%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	89.9%	89.8%	Y	Y	Y	N	Y	N	Y	N					Y	N			Y	Y	12	8	67%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

target) needed to meet AYP. In math, two elementary schools (Nemo and Island Grove) and one middle school (Pogesto) met the overall student population target with the confidence interval, although we know from Figure 3 that Nemo and Island Grove still failed to meet targets for some of their subgroups.

Figures 7 and 8 show the effect of confidence intervals on the reading proficiency rates for elementary and middle schools, respectively. One elementary school (Island Grove) and two middle schools (Pogesto and Filmore) met the overall reading targets through application of the confidence interval. **Overall, the application of the confidence interval has a moderate effect on whether sample schools met their overall targets in Indiana (or whether they make AYP).**[9]

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 through 8. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's complexity and length.

Table 3. Middle school subgroup performance of sample schools under the 2008 Indiana AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	61.7%	55.0%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	Y	Y	Y	18	5	28%	N	0
Barringer Charter	60.8%	60.7%	N	N	N	N			N	N	N	N			Y	Y					10	2	20%	N	0
ML Andrew	61.3%	59.0%	N	N	N	N			N	N	N	N			Y	N			Y	Y	12	3	25%	N	0
Pogesto	59.3%	64.8%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	63.7%	62.5%	N	N	N	N			N	N	N	N			N	N			Y	Y	12	2	17%	N	0
Tigerbear	70.6%	59.8%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	0
Chesterfield	75.3%	61.7%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	1
Filmore	75.7%	69.4%	Y	Y	N	N	Y	N	Y	N					Y	Y			Y	Y	12	8	67%	N	1
Barbanti	69.4%	65.2%	Y	N	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	78.6%	68.5%	Y	N	N	N	N	N	Y	N	N	N			N	N			Y	Y	14	4	29%	N	0
Hoyt	81.1%	71.4%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	2
Black Lake	82.9%	71.0%	Y	Y	N	N	N		Y	N	N	N	Y	Y	Y	Y			Y	Y	15	9	60%	N	0
Lake Joseph	79.3%	75.5%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	81.8%	73.9%	Y	Y	N	N	N	N	Y	N	Y	Y			N	N			Y	Y	14	7	50%	N	1
Ocean View	83.7%	82.4%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	83.7%	83.4%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	88.3%	80.3%	Y	Y	N	N			Y	N			Y	Y	N	N			Y	Y	12	7	58%	N	3
Chaucer	89.6%	88.1%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet, because it has fewer subgroups. These figures do not, however, indicate which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, LEP students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Indiana rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Indiana AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	13	10	13
Students with limited English proficiency	5	3	7
Low-income students	17	6	8
African-American students	6	1	4
Asian/Pacific Islander students	0	0	0
Hispanic students	9	4	7
American Indian/Alaska Native students	0	0	0
White students	17	0	1

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Indiana AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	9	8	8
Low-income students	17	7	14
African-American students	11	7	9
Asian/Pacific Islander students	4	0	0
Hispanic students	14	7	8
American Indian/Alaska Native students	1	1	0
White students	17	0	0

- Four elementary schools (Clarkson, Maryweather, Few and JFK) failed to meet reading targets for their overall school population and three of these elementary schools (Clarkson, Maryweather, and Few) also failed to meet targets in math.
- Four middle schools (McBeal, Barringer, ML Andrew, and McCord) failed to meet both reading and math targets for overall populations, and four middle schools (Tigerbear, Chesterfield, Barbanti, and Kekata) failed to meet overall reading targets.
- Three of the 16 failing elementary schools (Hissmore, Winchester, and Forest Lake) did not make AYP because of a single subgroup (SWDs).

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, the per-

Table 6. Comparisons between schools that did and didn't make AYP in Indiana, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	2	16	2	16
Average student body size	243	312	124	951
Average % low income	18	50	42	45
Average % nonwhite	25	43	27	46
Average performance†	5.61	0.68	0.40	-0.11
Average % growth‡	100	117	109	97
Average number of targets to meet	7	10	6	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

formance of SWDs is proving most challenging for schools under Indiana's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 30. In fact, all but three elementary schools and all of the middle schools in the study with qualifying SWD subgroups failed to make targets in math and all schools with such subgroups failed in reading. Students with LEP are also struggling to meet the state's targets; all schools with a large enough LEP population to qualify as a separate subgroup failed to meet reading targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Indiana's NCLB accountability system is, in most respects, behaving like those in other states. For example, Roosevelt and Wayne Fine Arts are among those schools that made AYP in the greatest number of states—28 and 21, respectively. And these schools made AYP in Indiana, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Indiana.

But Indiana is also home to a few anomalies. Consider Pogesto Middle School (Figure 4). Even with its relatively low average performance it made AYP in Indiana, but failed to do so in 13 of 28 states. Its AYP success in Indiana is likely attributable to the relatively small number of targets (four) it had to meet (as shown in Table 3). **In addition, Indiana has relatively easy proficiency standards, compared to other states, and a lenient confidence interval.** A second anomaly is apparent with Winchester Elementary, which made AYP in most of the states examined, but failed to make AYP in Indiana because of its SWD subgroup. This may be because Indiana uses a smaller minimum subgroup size than most other states, meaning that **schools in Indiana are accountable for more subgroups than similar schools in other states.**

This is consistent with the patterns shown in Table 6, which compares the sample schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that made AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus

fewer targets to meet), and much lower percentages of traditionally academically disadvantaged (e.g., low-income) students. Similarly, middle schools that made AYP have slightly higher performing students, on average, than middle schools that failed to make AYP, but have far smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

The study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Indiana's AYP rules (and AMOs) for 2008. We found that only 2 elementary schools and 2 middle schools—4 out of a sample of 36—made AYP in Indiana. Looking across the 28 state accountability systems examined in the study, this puts Indiana at the low end of the distribution in terms of the numbers of schools making AYP, as shown in Figure 1. **Though Indiana has relatively easy proficiency standards, it also uses fairly ambitious annual targets, and a smaller minimum subgroup size than most other states, meaning that schools in Indiana are accountable for more subgroups than similar schools in other states. All of these factors potentially inhibit the chances of a school making AYP in the Hoosier State.**

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Indiana's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Almost all of the sample schools made AYP in California for their student populations as a whole (i.e., without considering subgroup results). In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system[10] doesn't the massive failure of these students to meet Indiana's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

[10] See footnote 5.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Kansas

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Kansas's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Kansas's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Kansas's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Kansas's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **16 of 18 elementary schools** and **17 of 18 middle schools in our sample would fail to make adequate yearly progress** in 2008 under Kansas's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students. But it's also partially explained by Kansas's demanding annual targets for students (roughly 75% of students were expected to meet proficiency targets in 2008).
- **Looking across the 28 state accountability systems examined in the study, only two states passed fewer of the sample elementary schools than Kansas (Kansas ties 5 other states with only 2 elementary schools making AYP). In addition, Kansas is one of 6 states with a single passing middle school in the sample (see Figure 1).**
- Many of the schools in our sample that failed to make AYP in Kansas are meeting expected targets for their overall populations but failed because of the performance of individual subgroups, particularly students with disabilities (SWDs) and students with limited English proficiency (LEP).[2]

Under **Kansas's** accountability system, 16 of 18 elementary schools and 17 of 18 middle schools in our sample fail to make AYP in 2008. This places Kansas near the low end of the state distribution in terms of the number of schools making AYP. Kansas's definitions of proficiency generally ranked about average compared with the standards set by the other states. However, Kansas's annual targets in reading (the percentage of students in various subgroups that have to meet proficiency) are relatively difficult to achieve. Specifically, 75.6 percent of a given population in any school would have to be proficient on the state reading exam for the school to make AYP in 2008. In addition, Kansas's minimum subgroup size (30) is slightly lower than in many of the other states we examined. This means that more groups of students are held separately accountable than would be in many other states. In fact, every single school with a limited English proficient (LEP) or students with disabilities (SWD) subgroup failed to make AYP in Kansas.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Kansas Assessment Program.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

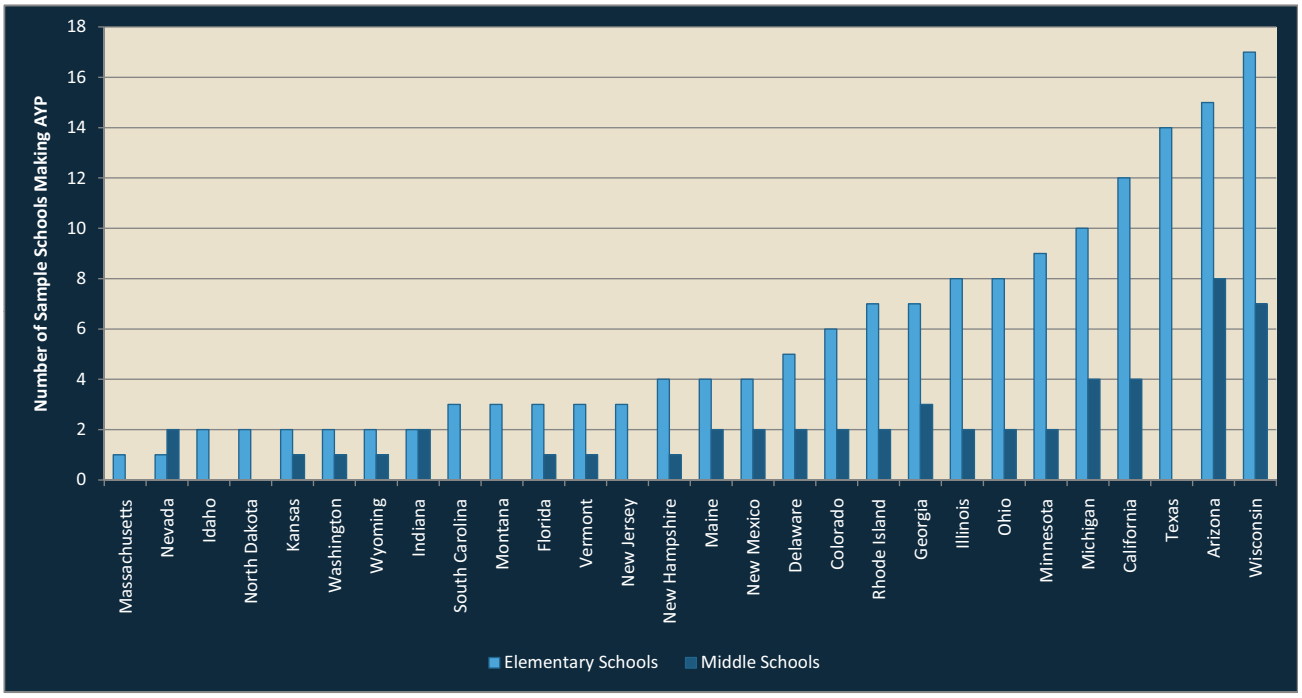

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- **In Kansas, as in most states, schools with fewer subgroups attained AYP more easily than schools with more subgroups**, even when their average student performance was much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As in other states, middle schools had greater difficulty reaching AYP in Kansas than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- Most states examined apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. However, **Kansas's 99% confidence interval give schools greater leniency than the more commonly used 95% confidence interval. Although the confidence interval did help a handful of schools in Kansas meet overall reading and math targets, it had little or no impact on final AYP outcomes because individual subgroups still failed to meet their targets (*all* of a school's subgroups must have met their targets for the school to make AYP)**.

- A strong predictor of whether or not a school will make AYP under Kansas's system is whether it has enough SWDs or English language learners[3] to qualify as a separate subgroup. Every single school with even one such subgroup failed to make AYP.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Kansas's tests and those of 25 other

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be slightly more negative than actual findings, mostly because of the differences in testing practices between the Measures of Academic Progress (MAP), the assessment we used in this study, and in the Kansas Assessment Program, the standardized state test. Specifically, the U.S. Department of Education has issued NCLB guidelines permitting schools to exclude small percentages of LEP students and SWDs from taking state tests, or providing them alternate assessments. In this study, however, no valid MAP scores were omitted from consideration.

states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies? NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African-American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

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

[6] We gave all schools in our sample pseudonyms in this report.

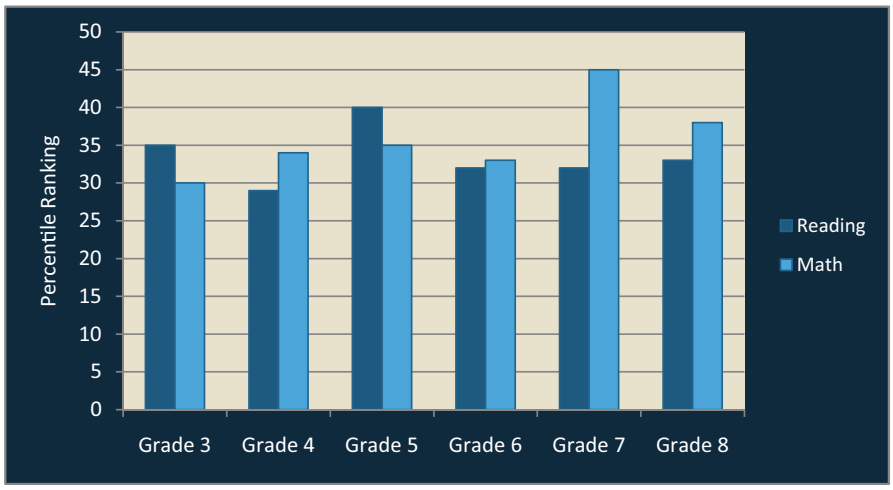

Figure 2. Kansas reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Kansas's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Kansas's cut scores are at or below the 45th percentile.

Proficiency cut score estimates for the Kansas Assessment System are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Kansas's definitions of proficiency generally ranked about average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Kansas test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Kansas AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Kansas's 2008

Table 1 shows the pertinent Kansas AYP rules that were applied to elementary and middle schools in this study. Kansas's minimum subgroup size is 30, which is slightly lower than in many of the other states we examined. This means that Kansas's schools would have to account for more subgroups than would similar schools in other states. Furthermore, although most states also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates, Kansas's 99% confidence interval gives schools greater leniency than the more commonly used 95% confidence interval. So for instance, while schools are supposed to get 75.6% of their grade 3-8 students to the "proficient" level on the state reading test, and 75.6% of the grade 3-8 students in each subgroup, applying the confidence interval means that the real target can be lower (particularly with smaller groups).[7]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

[7] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

Table 1. Kansas AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 30	
	Low-income students: 30	
	LEP students: 30	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	67.7	75.6
Grade 4	67.7	75.6
Grade 5	67.7	75.6
Grade 6	67.7	75.6
Grade 7	67.7	75.6
Grade 8	67.7	75.6
MATH		
Grade 3	62.5	73.4
Grade 4	62.5	73.4
Grade 5	62.5	73.4
Grade 6	62.5	73.4
Grade 7	62.5	73.4
Grade 8	62.5	73.4

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Kansas's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Kansas's 2008 AYP rules. **Only 2 elementary schools out of 16 made AYP**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The schools making AYP are in the right half of the figure, meaning that the highest performing students were found at these schools.

Yet almost without regard to average student performance, the only schools actually to make AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet. For example, Wayne Fine Arts passed, but had only eight targets – two in reading and math for the overall population, two in reading and math for its low-income population, two in reading and math for its Asian/Pacific Islander population, and two for its white population.

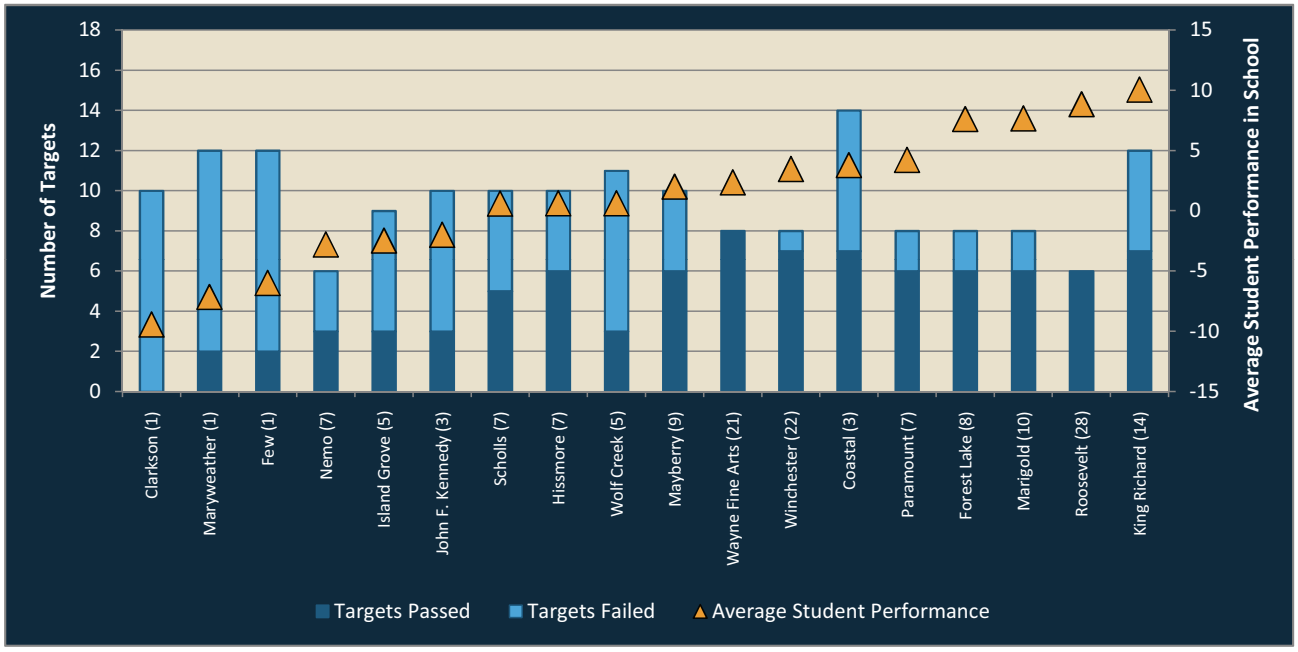

Figure 3. AYP performance of the elementary school sample under Kansas's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Kansas's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Winchester Elementary, for example, met seven of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Kansas AYP rules. **Of 18 middle schools in our sample, only one made AYP** (Walter Jones), which has relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' overall math proficiency rates are aided by Kansas's confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the AMO needed to meet AYP. Figures 5 and 6 show that four of the sample elementary schools (Nemo, Island Grove, JFK, and Wolf Creek) and three middle schools (Kekata, Hoyt, and Lake Joseph) are assisted by the confidence interval (note how the orange line falls within the light blue band). We know from Figures 3 and 4, however, that all of these schools failed to make AYP because of subgroup performance.

The effect of confidence intervals on the reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, four elementary schools (Hissmore, Mayberry, Coastal, and Paramount) and three middle schools (Pogesto, Hoyt, and Zeus) met the overall target with the confidence interval, although these schools still failed to meet all their subgroup targets (see Figures 3 and 4). **So, though the confidence interval does help some schools to meet overall reading and math targets, it has little or no impact on final AYP outcomes since individual subgroups failed to meet targets.**[8]

[8] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

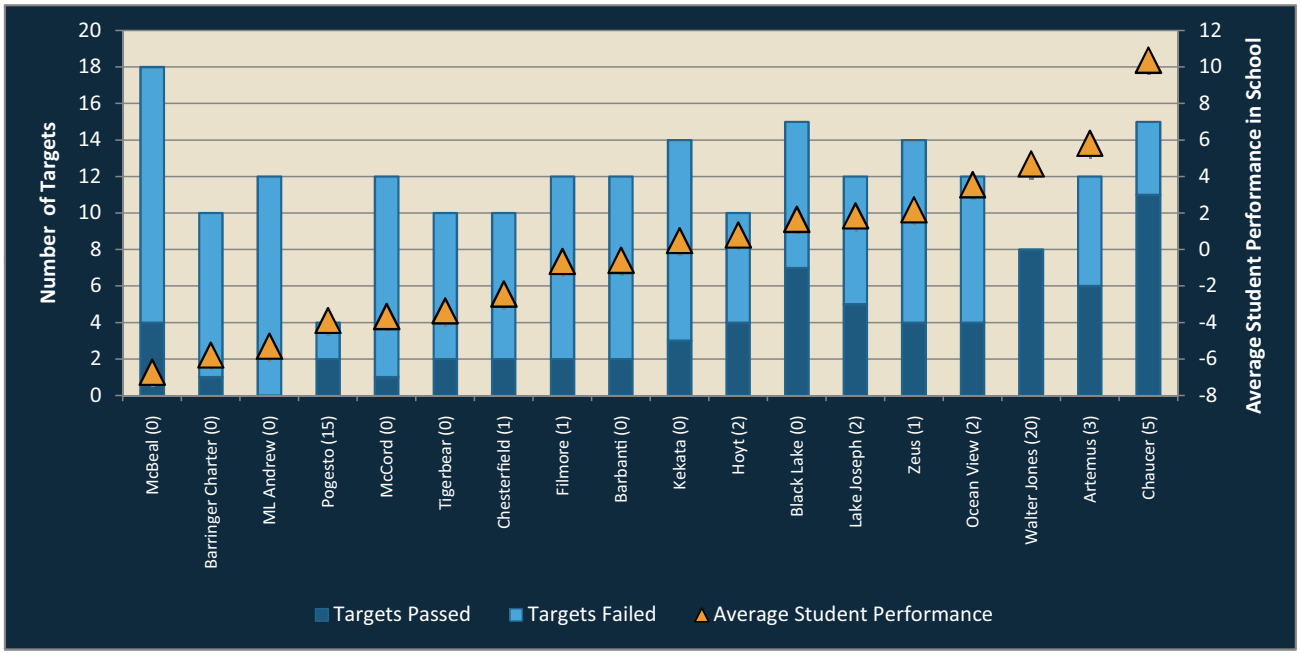

Figure 4. AYP performance of the middle school sample under Kansas's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Kansas's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Chaucer, for example, met 11 of its 15 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

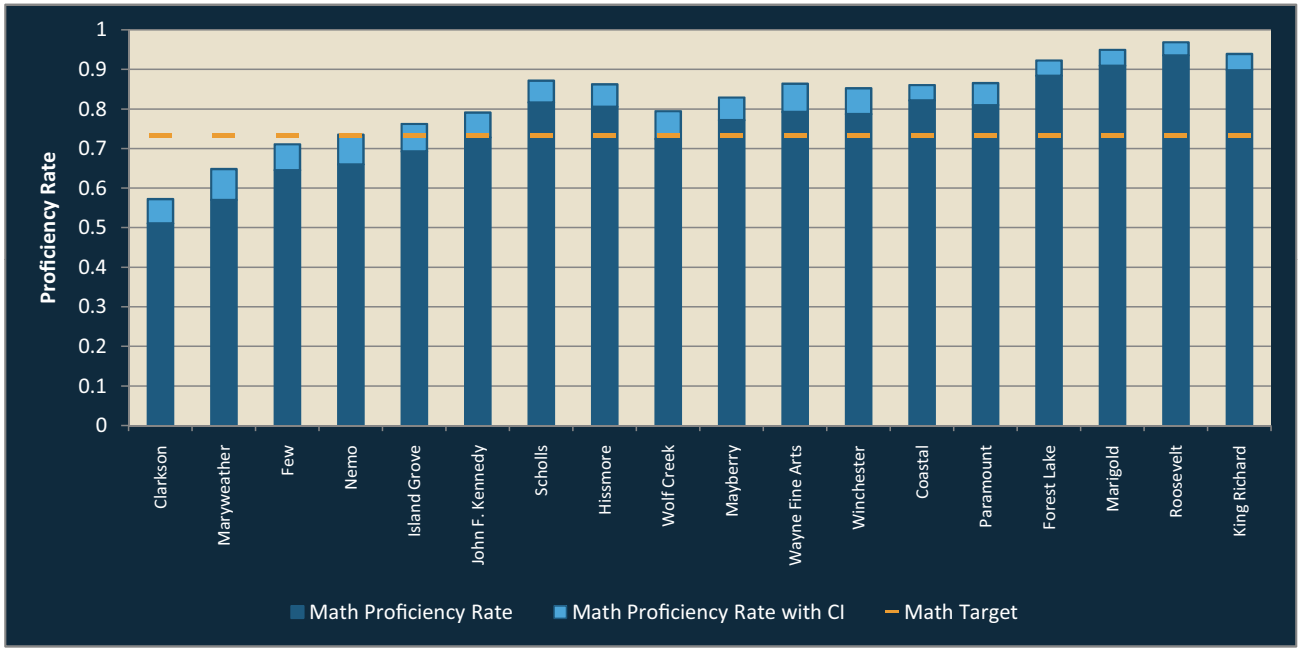

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that four schools (Nemo, Island Grove, JFK, and Wolf Creek) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

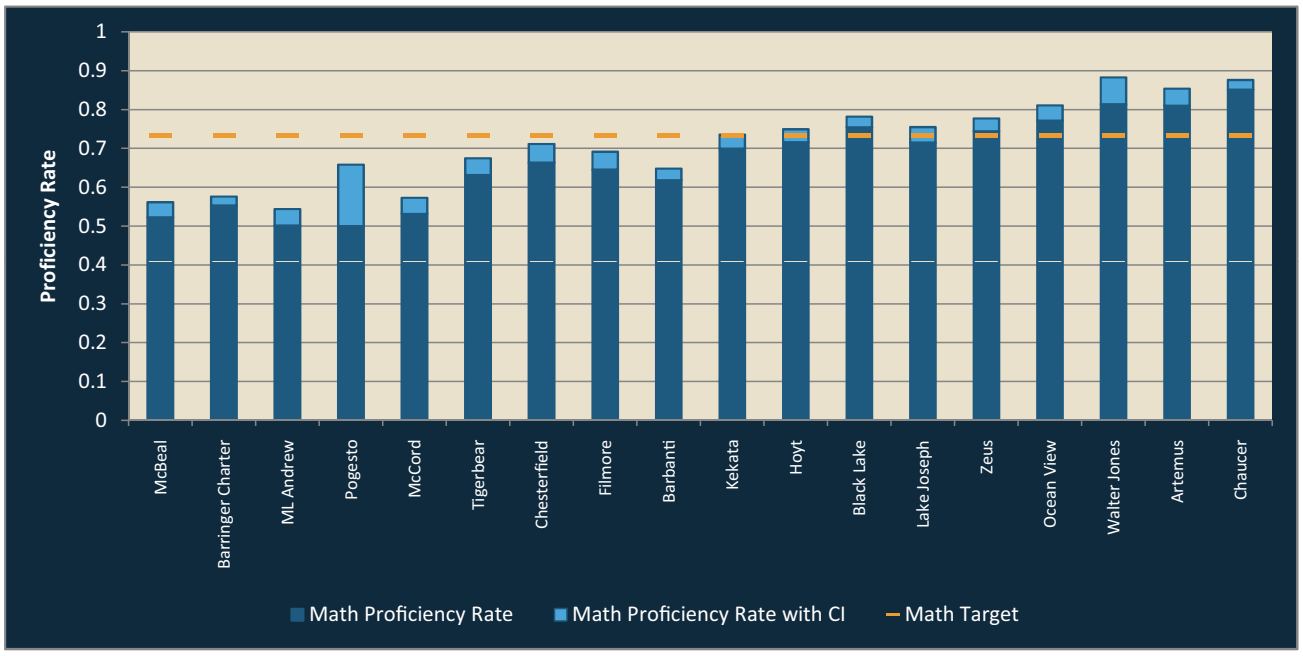

Figure 6. Impact of the confidence interval on middle school math proficiency rates under the 2008 Kansas AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that three schools (Kekata, Hoyt, and Lake Joseph) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP purposes include SWDs, LEP students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the Kansas rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Three elementary schools (Clarkson, Maryweather, and Few) failed to meet both math and reading targets for their overall school population.
- Five other elementary schools (Nemo, Island Grove, JFK, Scholls, and Wolf Creek) in the sample failed to meet their reading targets for their overall populations.
- Eight of the 17 failing middle schools in the sample (McBeal, Barringer, ML Andrew, McCord, Tigerbear, Chesterfield, Filmore, and Barbanti) failed for both reading and math for their overall populations.
- Most schools did not make AYP because of more than one subgroup.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. We see that the performance of SWDs is proving especially challenging under the Kansas accountability system. In fact, every SWD group at the middle

Table 2. Elementary school subgroup performance of sample schools under the 2008 Kansas AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	51.1%	36.1%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Maryweather	57.1%	47.0%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	1
Few	64.6%	48.5%	N	N	N	N	N	N	N	N					Y	N			Y	N	12	2	17%	N	1
Nemo	66.0%	63.7%	Y	N					N	N									Y	Y	6	3	50%	N	7
Island Grove	69.3%	65.4%	Y	N				N	N	N					N	N			Y	Y	9	3	33%	N	4
JFK	72.9%	57.5%	Y	N	N	N			Y	N	N	N							Y	N	10	3	30%	N	3
Scholls	81.7%	66.5%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	7
Hissmore	80.6%	69.8%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	7
Wolf Creek	72.5%	66.7%	Y	N	N	N		N	N	N					N	N			Y	Y	11	3	27%	N	5
Alice Mayberry	77.2%	71.3%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	9
Wayne Fine Arts	79.3%	77.6%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	78.8%	77.3%	Y	Y	Y	N									Y	Y			Y	Y	8	7	88%	N	22
Coastal	82.2%	72.9%	Y	Y	N	N	N	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	3
Paramount	81.0%	73.9%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	88.5%	83.3%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	91.0%	84.1%	Y	Y	Y	N			Y	N									Y	Y	8	6	75%	N	10
Roosevelt	93.6%	90.5%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	89.9%	86.4%	Y	Y	N	N	Y	N	Y	N					Y	N			Y	Y	12	7	58%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

school level failed to meet targets in both reading and math. A similar problem exists for students with limited English proficiency. All of those subgroups failed to meet their targets, save for one passing (in math) at the elementary level (King Richard).

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Kansas's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt and Wayne Fine Arts, made AYP in the greatest number of states—28 and 21, respectively. And these schools made AYP in Kansas, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Kansas.

But Kansas is also home to an anomaly. Winchester Elementary (see Figure 3) made AYP in 22 of the 28 states in our sample, but not in Kansas. In examining Table 2, we can see that Winchester missed only one target in reading for its SWD subgroup. This may be because

Table 3. Middle school subgroup performance of sample schools under the 2008 Kansas AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	52.3%	55.9%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	55.3%	57.2%	N	N	N	N			N	N	N	N			Y	N					10	1	10%	N	0
ML Andrew	50.1%	59.8%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Pogesto	50.0%	68.5%	N	Y															N	Y	4	2	50%	N	15
McCord Charter	53.2%	63.0%	N	N	N	N			N	N	N	N			N	N			N	Y	12	1	8%	N	0
Tigerbear	63.2%	61.0%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	0
Chesterfield	66.3%	63.0%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	1
Filmore	64.6%	71.1%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	1
Barbanti	61.8%	66.2%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Kekata	69.9%	69.0%	Y	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	3	21%	N	0
Hoyt	71.6%	72.4%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	2
Black Lake	75.4%	72.4%	Y	N	N	N	N		N	N	N	N	Y	Y	Y	Y			Y	Y	15	7	47%	N	0
Lake Joseph	71.5%	76.9%	Y	Y	N	N	N	N	N	Y					N	N			Y	Y	12	5	42%	N	2
Zeus	74.4%	74.4%	Y	Y	N	N	N	N	N	N	N	N			N	N			Y	Y	14	4	29%	N	1
Ocean View	77.2%	83.4%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	81.4%	81.1%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	81.0%	81.8%	Y	Y	N	N			N	N			Y	Y	N	N			Y	Y	12	6	50%	N	3
Chaucer	85.2%	88.1%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Kansas's minimum subgroup size is somewhat smaller than in most other states examined, meaning that school may have more accountable subgroups under Kansas rules than it would in other states.

This is consistent with the patterns shown in Table 6, which compares schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations (especially at the middle school level), fewer subgroups (and thus fewer targets to meet), and lower percentages of low-income and nonwhite students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Kansas's AYP rules (and AMOs) for 2008. We found that only 2 elementary schools and 1 middle school—3 out of a sample of 36— make AYP in Kansas. Looking across the 28 state accountability systems examined in the

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Kansas AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	13	11	13
Students with limited English proficiency	7	4	7
Low-income students	17	6	14
African-American students	6	1	5
Asian/Pacific Islander students	0	0	0
Hispanic students	9	4	8
American Indian/Alaska Native students	0	0	0
White students	17	0	2

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Kansas AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	9	9	8
Low-income students	17	15	14
African-American students	11	10	10
Asian/Pacific Islander students	4	0	0
Hispanic students	14	10	11
American Indian/Alaska Native students	1	1	1
White students	17	3	1

study, this puts Kansas at the low end of the sample distribution in terms of the number of schools making AYP (see Figure 1). Part of the reason that Kansas has so many schools not making AYP is that its annual targets are somewhat high (roughly 75% of students were expected to meet targets in 2008).

The overriding goal of the NCLB is to eliminate education disparities within and across states; it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Kansas's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or minority youngsters. Many of the sample schools met the Kansas reading and math targets for their student populations as

Table 6. Comparisons between schools that did and didn't make AYP in Kansas, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	2	16	1	17
Average student body size	243	312	165	900
Average % low income	18	50	38	45
Average % nonwhite	25	43	33	45
Average performance†	5.61	0.68	4.69	-0.33
Average % growth‡	100	117	111	97
Average number of targets to meet	7	10	8	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

a whole, that is, without considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their students weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of the student population has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[9] doesn't the massive failure of middle school students to meet Kansas's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other pupils, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP

[9] See footnote 4.

scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Maine

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines the NCLB accountability system in Maine—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough the Maine system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under the Maine system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against the Maine AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **14 of 18 elementary schools** and **16 of 18 middle schools** in our sample failed to make adequate yearly progress in 2008 under Maine's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students. It's also partly explained by Maine's proficiency cut scores which are above average, or relatively difficult, compared with the standards set by the other states in the study. In addition, Maine's minimum subgroup size is 20, which is quite small compared to most other states. This means that schools in Maine will have more accountable subgroups than similar schools in other states.
- Looking across the 28 state accountability systems examined in the study, **12 states passed more of the sample elementary schools than did Maine, while 13 states passed fewer elementary schools. In other words, Maine was about in the middle** (see Figure 1).
- Nearly all of the schools in our sample that failed to make AYP in Maine are meeting expected targets for their overall populations but failing because of the performance of individual subgroups.[2]
- As is the case in other states, schools with fewer subgroups attain AYP more easily in Maine than schools with more subgroups, even when their average student performance is much lower. In other words,

Maine's AYP rules place the state toward the mid to lower end of the state distribution in terms of the number of schools making AYP. Maine's proficiency cut scores generally ranked above average, or relatively difficult, compared with the standards set by the other states in the study. In addition, Maine's minimum subgroup size is 20, which is quite small compared to most other states. This means that more subgroups are held accountable in Maine than would be in other states. In fact, all but two schools with limited English proficient (LEP) or students-with-disabilities (SWD) subgroups failed to make AYP, in part because these students did not meet the state's proficiency targets in math and reading. Students with disabilities had a particularly hard time meeting their AYP targets at the middle school level.

[1] A cut score is the minimum score a student must receive on NEWA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Maine Education Assessment (MEA).

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

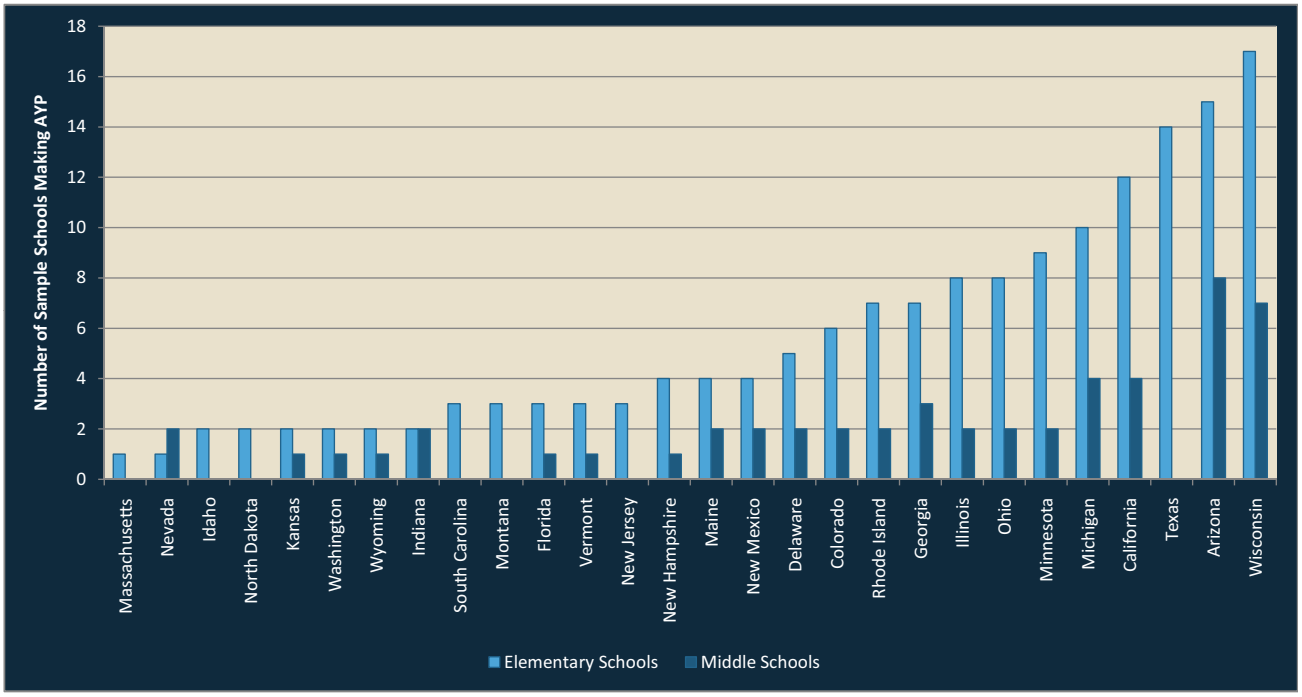

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

schools with greater diversity and size face greater challenges in making AYP.

- Middle schools have greater difficulty reaching AYP in Maine than do elementary schools, primarily because their student populations are larger and therefore, have more qualifying subgroups—not because their student achievement is any lower than in the elementary schools.

- A strong predictor of whether or not a school will make AYP under the Maine system is whether it has enough students with disabilities (SWD) or English language learners to qualify as a separate subgroup. Nearly all schools with limited English proficient (LEP)[3] or SWD subgroups failed to make AYP, in part because these students did not meet the state's proficiency targets in reading.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Maine's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency stan-

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Maine Education Assessment, the standardized state test. 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.

dards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Maine Education Assessment (MEA) are taken from *The Proficiency Illusion* (as shown in Figure 1), which found that **Maine's proficiency cut scores were generally ranked above average, or relatively difficult, compared with the standards set by the other 25 states in that study**. These cut scores were used to estimate whether students would have scored as proficient or better on the Maine test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Maine AYP rules for 2008. In other words,

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

[6] We gave all schools in our sample pseudonyms in this report.

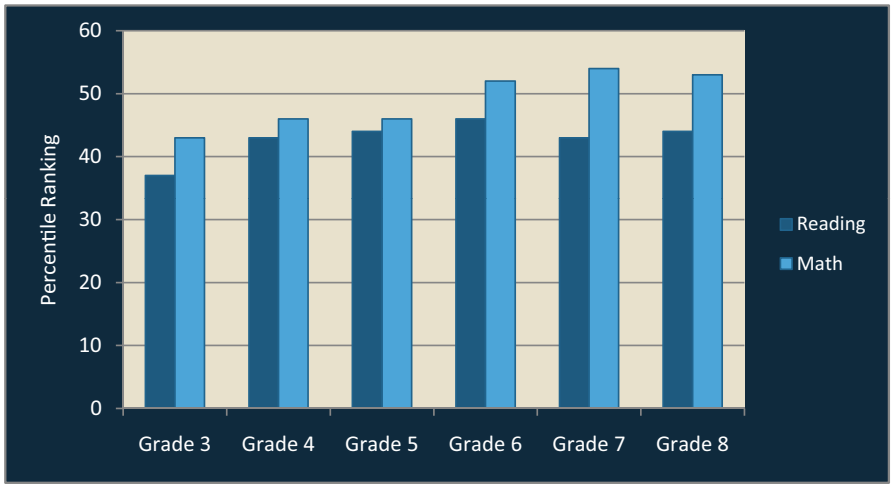

Figure 2. Maine reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Maine's cut scores (as proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Maine's cut scores are below the 55th percentile.

Table 1. Maine AYP Rules for 2008

Subgroup minimum *n*	Race/ethnicity: 20	
	SWDs: 20	
	Low-income students: 20	
	LEP students: 20	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	49
Grade 4	34	49
Grade 5	n/a	49
Grade 6	n/a	50
Grade 7	n/a	50
Grade 8	35	50
MATH		
Grade 3	n/a	32
Grade 4	12	32
Grade 5	n/a	32
Grade 6	n/a	33
Grade 7	n/a	33
Grade 8	13	33

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not available

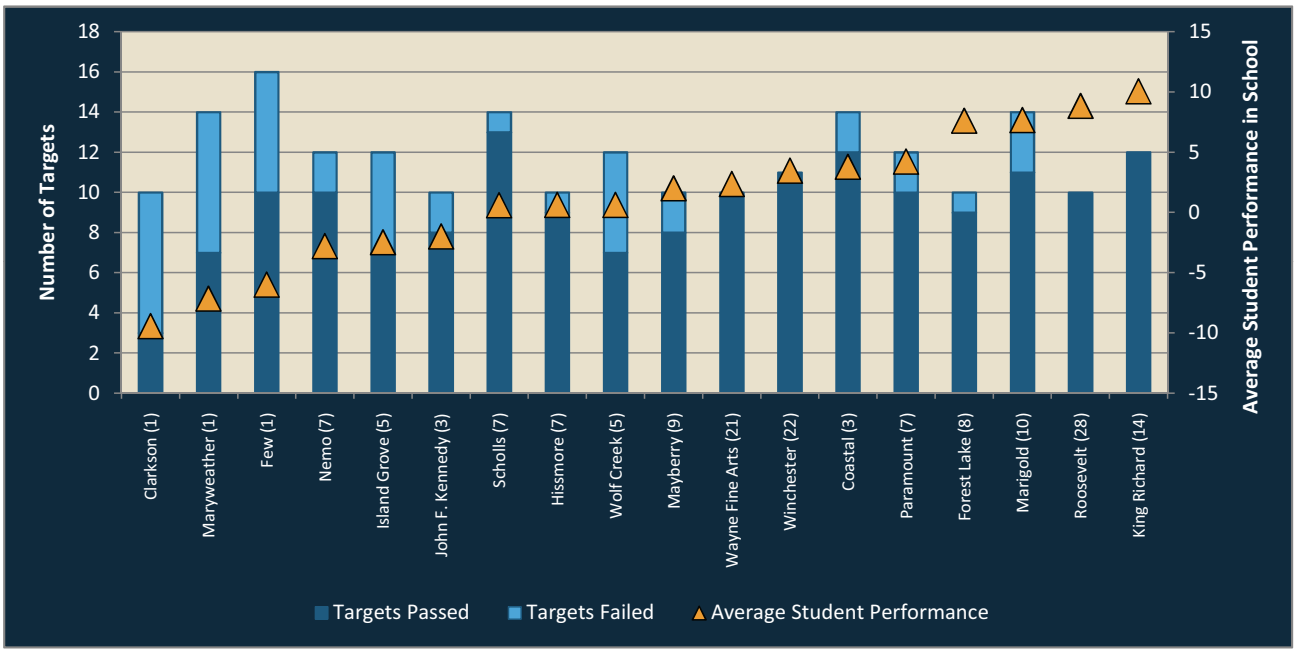

Figure 3. AYP performance of the elementary school sample under Maine's 2008 AYP Rules

Note: This figure indicates how each of the elementary schools within the sample fared under the Maine AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Mayberry, for example, met eight of its ten targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

the school data and our proficiency cut score estimates are from 2005–2006, but we are applying them against the Maine 2008 AYP rules.

Table 1 shows the pertinent Maine AYP rules that were applied to elementary and middle schools in the current study. Maine's minimum subgroup size is 20, which is quite small compared to most other states examined in the study. This means that schools in Maine will have more accountable subgroups than similar schools in other states. Maine, like the majority of states examined in the study, applies the 95% confidence intervals to their measurements of student proficiency rates, which makes it easier to achieve their annual measurable objectives. So, for instance, even though schools are supposed to get 50% of their grade 6 students to the proficient level on the state reading test, as well as 50% of the grade 6 students in each subgroup, applying the confidence interval means that the real target can be lower, particularly with smaller groups.

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test-participation rates are beyond the scope of the study. Most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

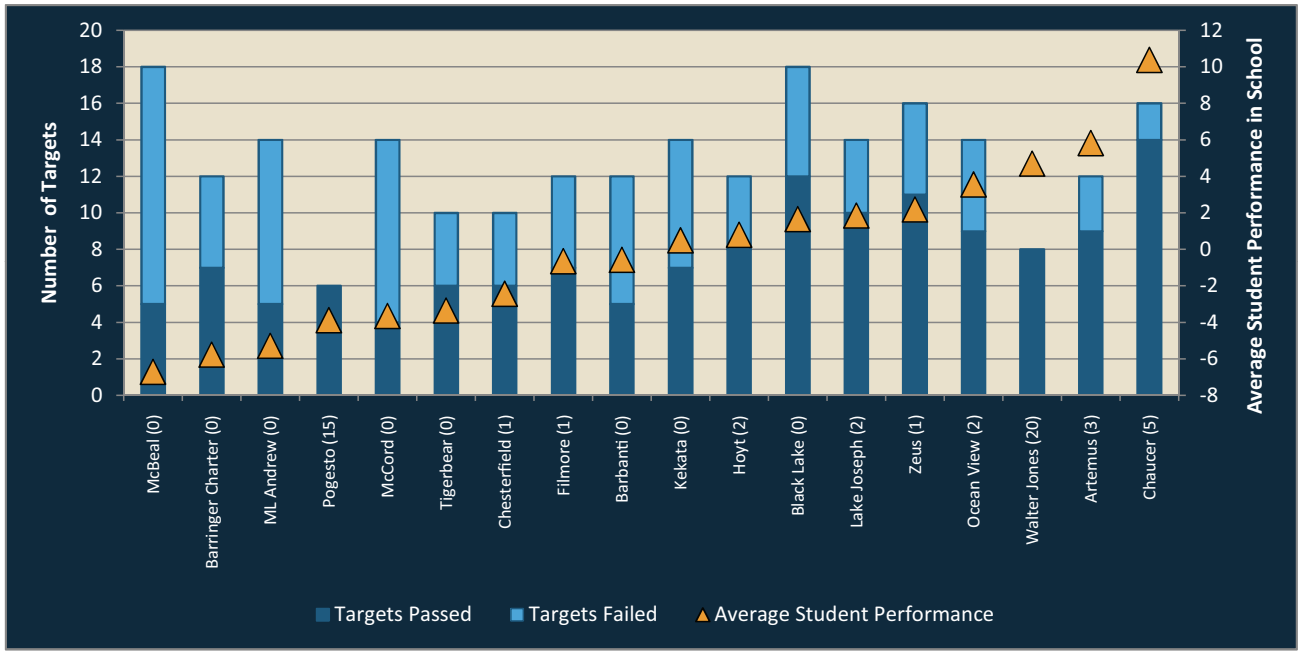

Figure 4. AYP performance of the middle school sample under Maine's 2008 AYP Rules

NNote: This figure shows how each of the middle schools within the sample fared under the AYP rules in Maine (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Artemus Middle School, for example, met nine of its twelve targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

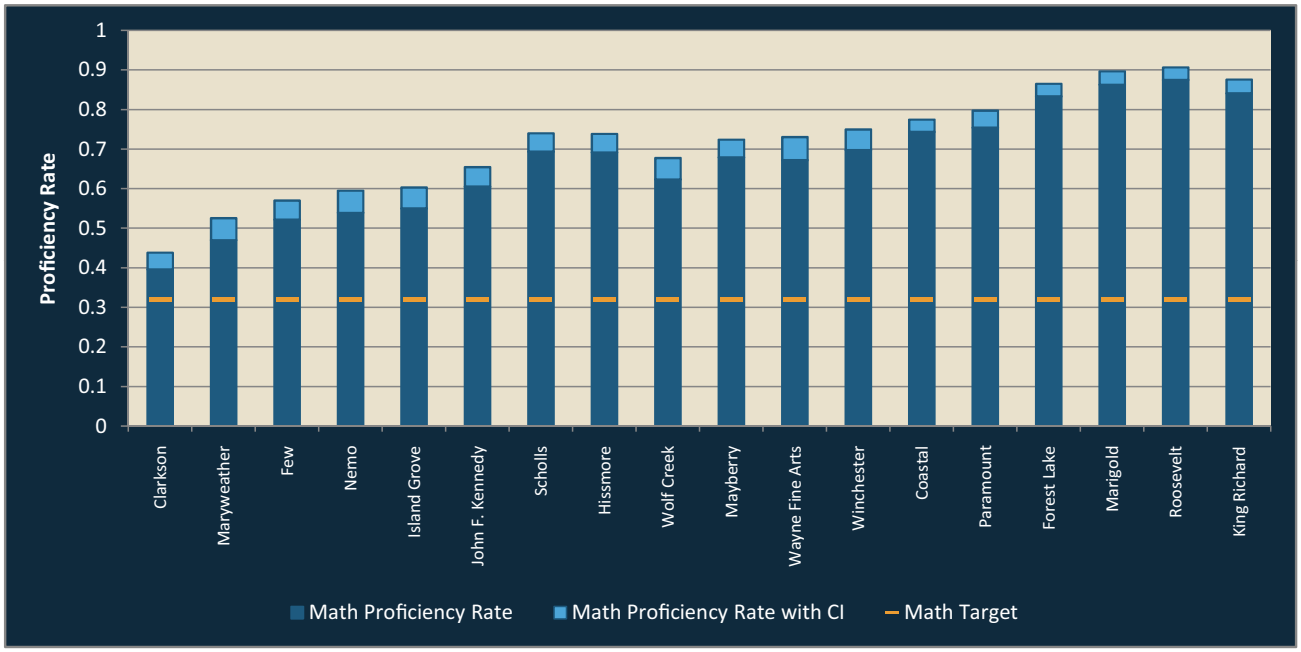

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under Maine's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

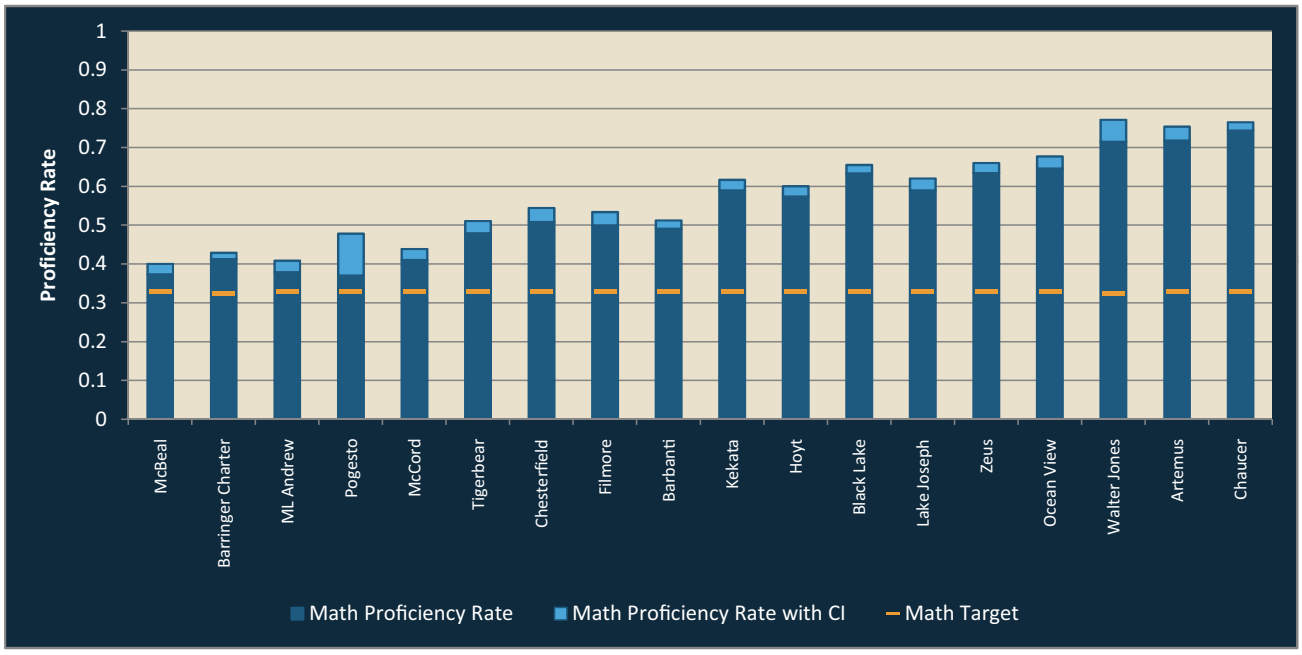

Figure 6. Impact of the confidence interval on middle school math proficiency rates under Maine's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample middle schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under Maine's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Maine's 2008 AYP rules. **Only 4 schools (Wayne Fine Arts, Winchester, Roosevelt, and King Richard) made AYP while 14 failed to make it**. The triangles in the Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. All schools making AYP are in the right half of the figure, meaning that these schools contain the highest performing students.

Yet almost without regard to average student performance, the only schools actually to make AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet. For example, Wayne Fine Arts made AYP, but only has ten targets.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Maine AYP rules. **Out of 18 in our sample, only 2 middle schools make AYP** – 1 low-performance school (Pogesto), and 1 high-performance school (Walter Jones), both of which have few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' overall math proficiency rates are aided by Maine's confidence interval for elementary and middle schools, respectively. On this figure, the darker portions of the bars show the actual proficiency rates at each school, and the lighter portions of the bars show the degree to which these proficiency rates were increased by the applying the confidence interval. The orange lines show the AMO needed to meet AYP. **These figures show that none of the sample elementary or middle schools were assisted**

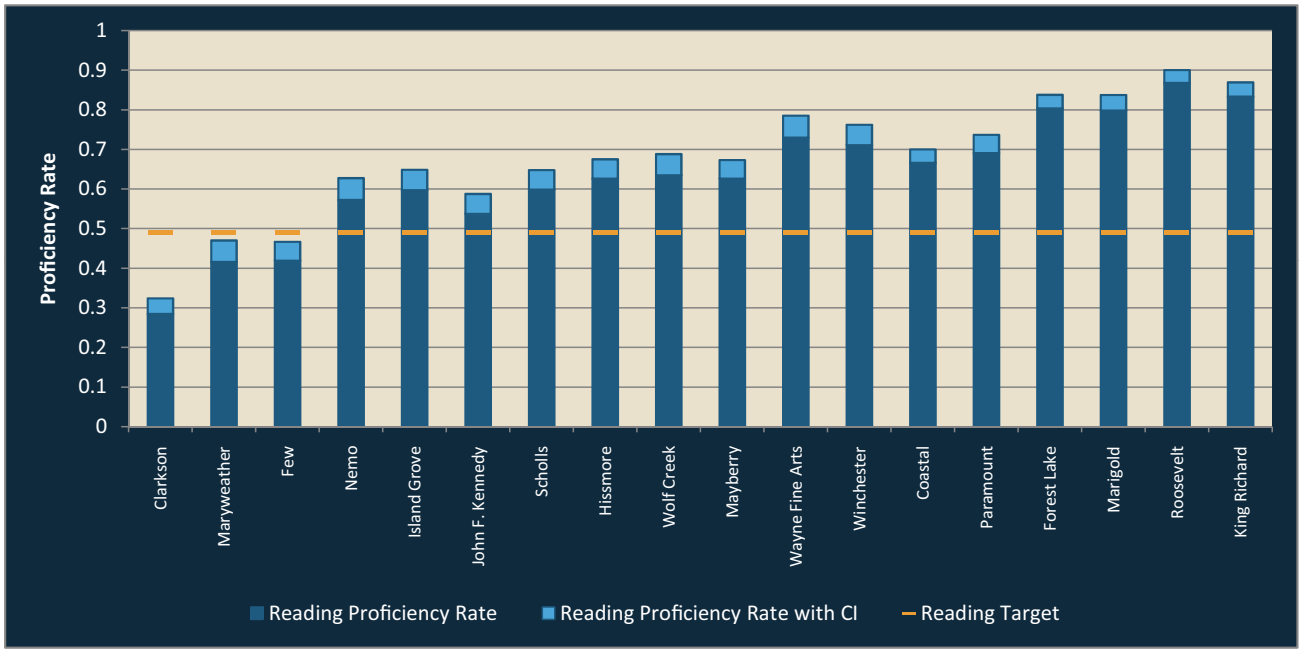

Figure 7. Impact of the confidence interval on elementary school reading proficiency rates under Maine's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

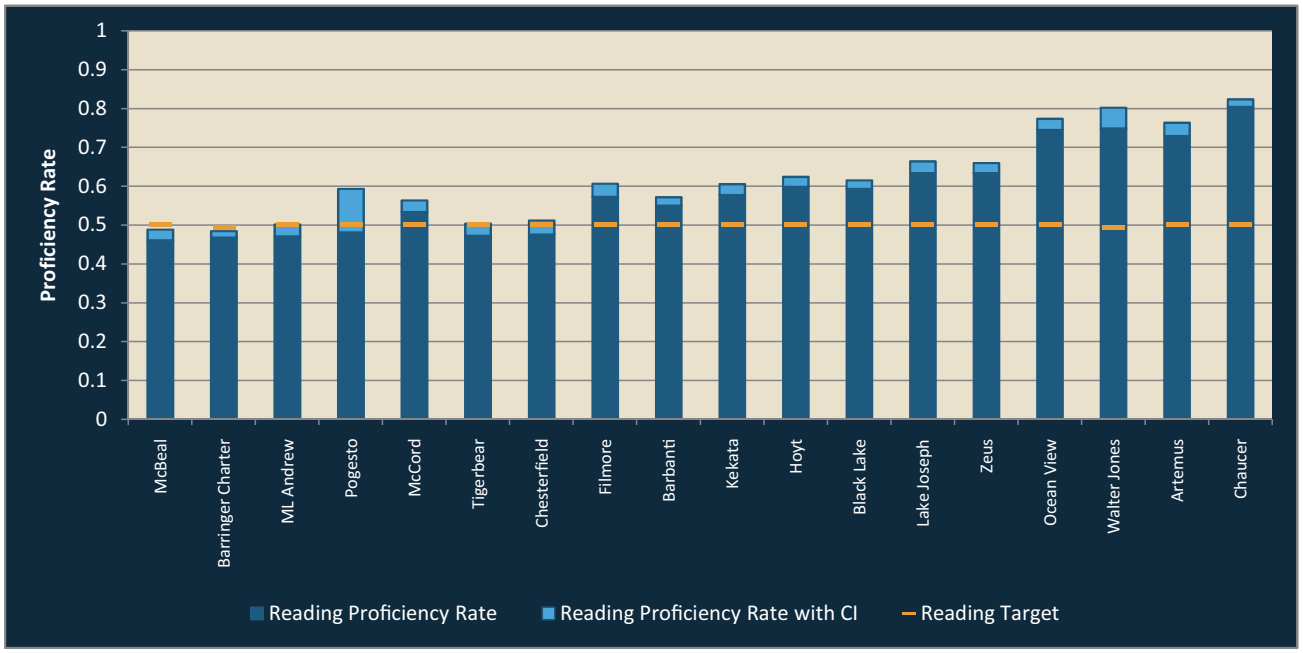

Figure 8. Impact of the confidence interval on middle school reading proficiency rates under Maine's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample middle schools (Pogesto and Chesterfield) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Maine AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	39.6%	28.5%	Y	N	N	N	N	N	Y	N					Y	N					10	3	30%	N	1
Maryweather	47.0%	41.6%	Y	N	N	N	Y	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	1
Few	52.2%	41.9%	Y	N	Y	N	Y	N	Y	N	Y	Y			Y	N	Y	N	Y	Y	16	10	63%	N	1
Nemo	54.0%	57.2%	Y	Y	Y	N			Y	N	Y	Y			Y	Y			Y	Y	12	10	83%	N	7
Island Grove	55.0%	59.7%	Y	Y	N	N	N	N	Y	Y					Y	N			Y	Y	12	7	58%	N	4
JFK	60.6%	53.7%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	3
Scholls	69.4%	59.9%	Y	Y	Y	N	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	13	93%	N	7
Hissmore	69.2%	62.6%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	7
Wolf Creek	62.4%	63.5%	Y	Y	N	N	Y	N	Y	N					Y	N			Y	Y	12	7	58%	N	5
Alice Mayberry	67.9%	62.6%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	9
Wayne Fine Arts	67.2%	73.0%	Y	Y					Y	Y	Y	Y			Y	Y			Y	Y	10	10	100%	Y	21
Winchester	69.8%	71.1%	Y	Y	Y	Y			Y	Y				Y	Y	Y			Y	Y	11	11	100%	Y	22
Coastal	74.4%	66.6%	Y	Y	Y	N	Y	N	Y	Y	Y	Y			Y	Y			Y	Y	14	12	86%	N	3
Paramount	75.5%	69.0%	Y	Y	Y	Y	N	N	Y	Y					Y	Y			Y	Y	12	10	83%	N	7
Forest Lake	83.4%	80.4%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	8
Marigold	86.3%	79.8%	Y	Y	Y	N	Y	N	Y	Y			Y	Y	Y	N			Y	Y	14	11	79%	N	10
Roosevelt	87.5%	86.8%	Y	Y					Y	Y	Y	Y			Y	Y			Y	Y	10	10	100%	Y	28
King Richard	84.1%	83.3%	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y			Y	Y	12	12	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

by the confidence intervals, because the math targets in Maine are low, relative to the schools' overall performance. In other words, the sample schools met the targets without the assistance of the confidence interval.

The effect of confidence intervals on reading proficiency rates for elementary and middle schools is similar (Figures 7 and 8). In reading, none of the elementary schools make use of the confidence interval to meet the overall target. Two of the sample middle schools (Pogesto and Chesterfield) met the overall target with the help of the confidence interval (see Figure 8), but we know that Chesterfield still failed to meet all its subgroup targets (Figure 4). **In short, the application of the confidence interval has only modest impact on AYP decisions for the sample schools in Maine.**[7]

[7] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figures 5 through 8. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 3. Middle school subgroup performance of sample schools under the 2008 Maine AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	37.4%	46.1%	Y	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	5	28%	N	0
Barringer Charter	41.3%	46.8%	Y	N	N	N			Y	N	Y	N			Y	Y			Y	Y	12	7	58%	N	0
ML Andrew	37.9%	47.1%	Y	Y	N	N	N	N	N	N	N	N			Y	N			Y	Y	14	5	36%	N	0
Pogesto	37.0%	48.1%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	15
McCord Charter	41.0%	53.4%	Y	Y	N	N	N	N	N	N	N	N			N	N			Y	Y	14	4	29%	N	0
Tigerbear	47.9%	47.2%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	0
Chesterfield	50.9%	47.6%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	1
Filmore	50.0%	57.3%	Y	Y	N	N	N	N	Y	Y					Y	N			Y	Y	12	7	58%	N	1
Barbanti	49.1%	55.0%	Y	Y	N	N	N	N	N	N					Y	N			Y	Y	12	5	42%	N	0
Kekata	58.9%	57.8%	Y	Y	N	N	N	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	0
Hoyt	57.4%	59.9%	Y	Y	N	N			Y	Y	Y	N			Y	N			Y	Y	12	8	67%	N	2
Black Lake	63.4%	59.3%	Y	Y	N	N	Y	N	Y	N	Y	N	Y	Y	Y	N	Y	Y	Y	Y	18	12	67%	N	0
Lake Joseph	59.0%	63.4%	Y	Y	N	N	N	N	Y	Y	Y	Y			Y	Y			Y	Y	14	10	71%	N	2
Zeus	63.5%	63.4%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	N			Y	Y	16	11	69%	N	1
Ocean View	64.6%	74.5%	Y	Y	Y	N	N	N	Y	N			Y	Y	Y	N			Y	Y	14	9	64%	N	2
Walter Jones	71.5%	74.9%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	71.8%	72.9%	Y	Y	Y	N			Y	N			Y	Y	Y	N			Y	Y	12	9	75%	N	3
Chaucer	74.4%	80.4%	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	14	88%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Where do schools fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet, thanks to fewer subgroups. These figures do not, however, indicate which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Maine rules, and the total number of states within the study in which that school met AYP.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Maine AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	5	13
Students with limited English proficiency	10	3	8
Low-income students	18	0	5
African-American students	11	0	2
Asian/Pacific Islander students	1	0	0
Hispanic students	14	0	6
American Indian/Alaska Native students	1	0	1
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Maine AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	13	16
Students with limited English proficiency	11	9	11
Low-income students	18	4	11
African-American students	12	3	9
Asian/Pacific Islander students	6	0	0
Hispanic students	15	2	11
American Indian/Alaska Native students	2	1	1
White students	18	0	0

The school-by-school findings in Tables 2 and 3 show that:

- Three elementary schools (Clarkson, Maryweather, and Few) and two middle schools (McBeal and Barringer) failed to meet the reading targets for their overall school population.
- No school failed to meet their overall targets in math.
- Four of the fourteen failing elementary schools (Scholls, Hissmore, Alice Mayberry, and Forest Lake) missed only for the SWD subgroup.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. First, elementary students did better in math than reading, perhaps because Maine's proficiency targets are lower in math than in reading at the elementary grades

(32% and 49%, respectively, as shown in Table 1). The performance of SWD students is also proving challenging for schools under Maine's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 20. In fact, all but two elementary and all middle schools in the study with a qualifying SWD subgroup failed to make AYP. Students with LEP are also struggling to meet the state's targets; all but two elementary schools with a large enough LEP population to qualify as a separate subgroup failed to meet their reading targets for these students.

A close look at Figures 3 and 4 indicates that Maine's NCLB accountability system is, in many respects, behaving similarly to those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all make AYP in the greatest number of states—28, 22, and 21, respectively. And these schools make AYP in Maine, too. Likewise, the elementary and middle schools that fail to make AYP in the greatest number of states also fail in Maine.

Other state reports contain a section comparing some of the characteristics of the sample schools that made AYP versus those that did not. In Maine, there were no striking differences between schools that made and didn't make AYP, other than the (expected) finding that the former had students with higher average student performance than the latter, as measured by NWEA reading and math tests.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would have fared under the Maine AYP rules (and AMOs) for 2008. We found that only 4 elementary schools and 2 middle schools–6 in all from a sample of 36–would have made AYP in Maine. Looking across the 28 state accountability systems examined in the study, this puts Maine in the middle of the distribution in terms of the number of schools making AYP (as shown in Figure 1).

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, the NCLB accountability system in Maine is working exactly as Congress intended: identifying as needing attention those schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Almost all the sample schools met the Maine AMO targets for their student populations as a whole, i.e., not considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils were not meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[8] does the massive failure of middle school students to meet Maine's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

[8] See footnote 4.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Massachusetts

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Massachusetts's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Massachusetts's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Massachusetts's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Massachusetts's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **17 of 18 elementary schools** and **all 18 middle schools** in our sample **failed to make AYP** in 2008 under Massachusetts's accountability system. (This very high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, **we find that virtually all the states (with the exception of Nevada, which ties Massachusetts) exceed Massachusetts in terms of the number of elementary schools making AYP. In addition, Massachusetts is one of only five states (along with Idaho, Montana, South Carolina, and North Dakota) that had no passing middle schools in the sample (see Figure 1).**[2]
- Middle schools had even greater difficulty reaching AYP in Massachusetts than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is any lower than in the elementary schools.
- The only school in Massachusetts that made AYP had only one subgroup (white).

There are several factors in **Massachusetts** which contribute to only one school making AYP in the study. First, the math proficiency standard ranges from a high of the 77th percentile in grade 4 to the 68th percentile in grades 6 and 8. This means that to be considered proficient, grade 4 students must perform better than 77% of all other students in the nation (calculated from the NWEA norms). The reading standard is somewhat lower, ranging from the 65th percentile in grade 4 to the 30th percentile in grade 8. Second, despite the fact that it's lower, Massachusetts still expects a high percentage (roughly 85%) of its grade 3-8 students to reach the reading standard in 2008. These two dynamics, combined with the fact that Massachusetts does not apply a confidence interval (margin of error) to proficiency rate calculations, contribute to only one school making AYP in the study.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Massachusetts Comprehensive Assessment System (MCAS).

[2] At the same time, it's important to note that Massachusetts has improved more than almost every state on the National Assessment of Educational Progress (NAEP) test. In 2007, for instance, it scored first in the nation in fourth- and eighth-grade math and reading.

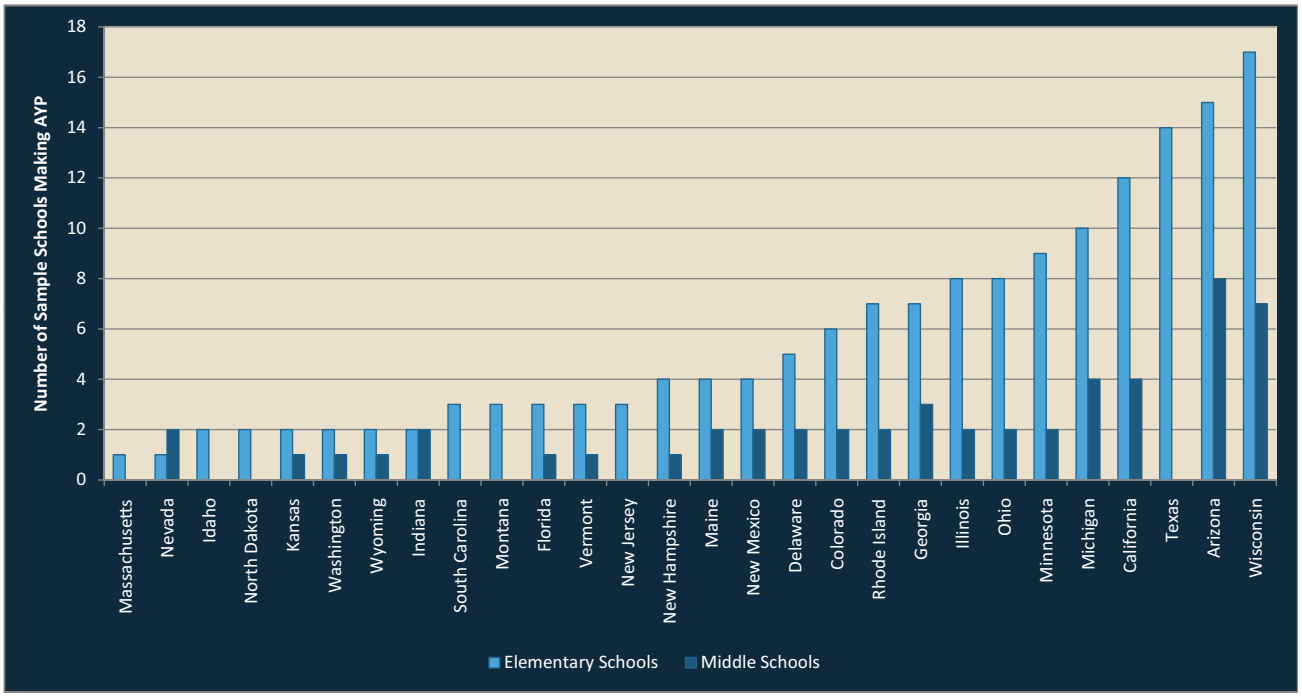

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- **Massachusetts's high proficiency standards mean that schools will have increasing difficulty in meeting the 100% proficiency requirements of NCLB by 2014.**

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Massachusetts's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as

low income[3] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states, but not Massachusetts, have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners,[4] among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[5]

Proficiency cut score estimates for the Massachusetts Comprehensive Assessment System (MCAS) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that **Massachusetts's definitions of proficiency generally ranked far above the average set by the other 25 states in that study**. These cut score were used to estimate whether students would have scored as proficient or better on the Massachusetts test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Massachusetts AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Massachusetts's 2008 AYP rules.

Table 1 shows the pertinent Massachusetts AYP rules that were applied to elementary and middle schools in the current study. Massachusetts's minimum subgroup size is 40, as long as that constitutes at least 5% of the student population; subgroups can't be larger than 200 students. The sliding minimum subgroup number used by Massachusetts is not used by most other states, but it means that for many schools, the actual minimum number will be larger than 40.[6]

Massachusetts, unlike most other states examined, does not apply a confidence interval (or margin of statistical

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

[4] Note that we use "students with limited English proficiency (LEP)" or "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup.

[5] We gave all schools in our sample pseudonyms in this report.

[6] This means that a school with a total population of 1000 would have a minimum subgroup size of 50 (i.e., 5%), but a school with only 200 students would have a minimum subgroup size of 40, since 5% of 200 (i.e., 10) is below the subgroup minimum of 40. Similarly, a hypothetical school of 5,000 would have a minimum subgroup size of 200, since 5% of 5,000 (i.e., 250) is greater than the subgroup maximum of 200.

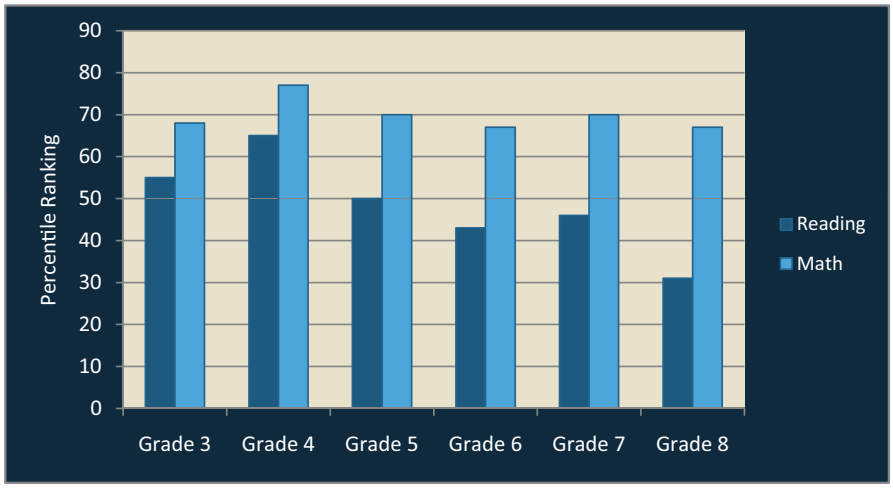

Figure 2. Massachusetts reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Massachusetts's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. Though Massachusetts's cut scores vary by grade and subject, all of the math cut scores and half of the reading cut scores are at or above the 50th percentile.

Table 1. Massachusetts AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 5% of the student population but with a minimum of 40 and maximum of 200	
	SWDs: 5% of the student population but with a minimum of 40 and maximum of 200	
	Low-income students: 5% of the student population but with a minimum of 40 and maximum of 200	
	LEP students: 5% of the student population but with a minimum of 40 and maximum of 200	
CI	**Applied to proficiency rate calculations?**	
	Not used	
AMOs	**Baseline proficiency levels as of 2002 (index)**	**2008 targets (index)**
READING/LANGUAGE ARTS		
Grade 3	70.7	85.4
Grade 4	70.7	85.4
Grade 5	n/a	85.4
Grade 6	n/a	85.4
Grade 7	70.7	85.4
Grade 8	n/a	85.4
MATH		
Grade 3	n/a	76.5
Grade 4	53.0	76.5
Grade 5	n/a	76.5
Grade 6	53.0	76.5
Grade 7	n/a	76.5
Grade 8	53.0	76.5

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).
Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not available

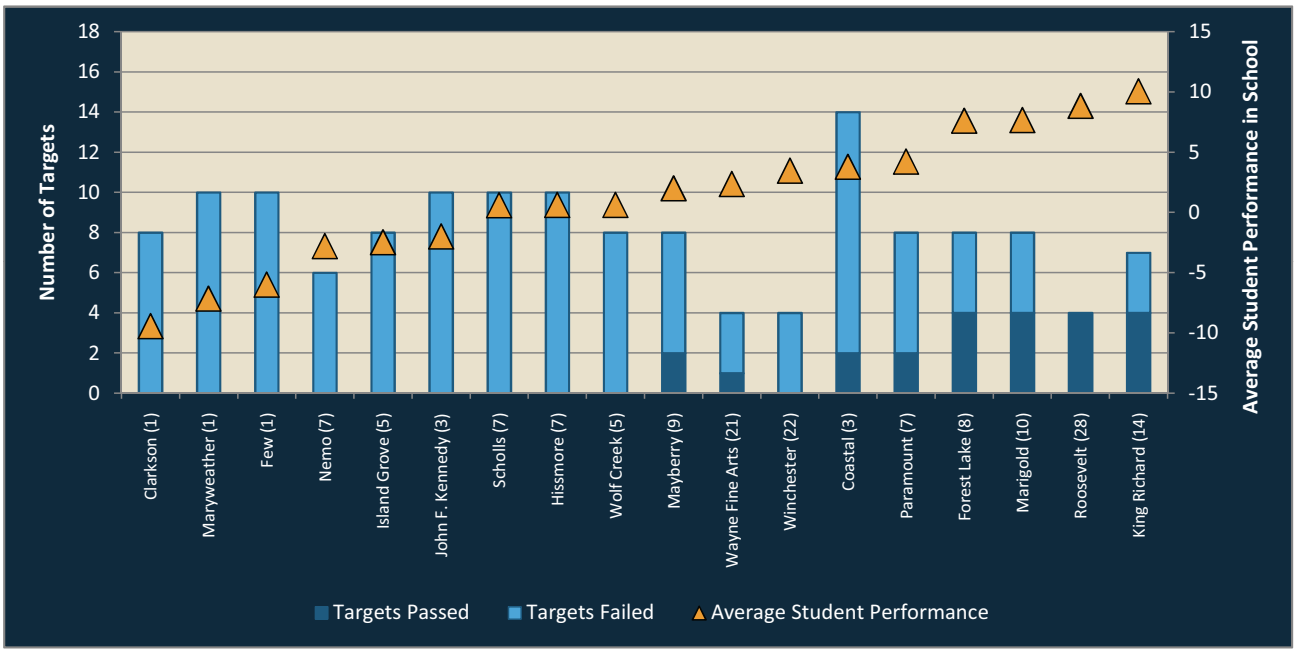

Figure 3. AYP performance of the elementary school sample under Massachusetts's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Massachusetts's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Marigold Elementary, for example, met four of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

error) to measurements of student proficiency rates. This means that **schools in Massachusetts will have a more difficult time meeting their proficiency targets than similar schools in other states that do use confidence intervals**. Unlike most states examined, however, Massachusetts targets are measured against an index rather than a proficiency percentage, meaning that partially proficient students receive partial credit.[7]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are

[7] Six of the states (Minnesota, Rhode Island, Vermont, Wisconsin, New Hampshire, as well as Massachusetts) in our 28-state sample use an index that 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.

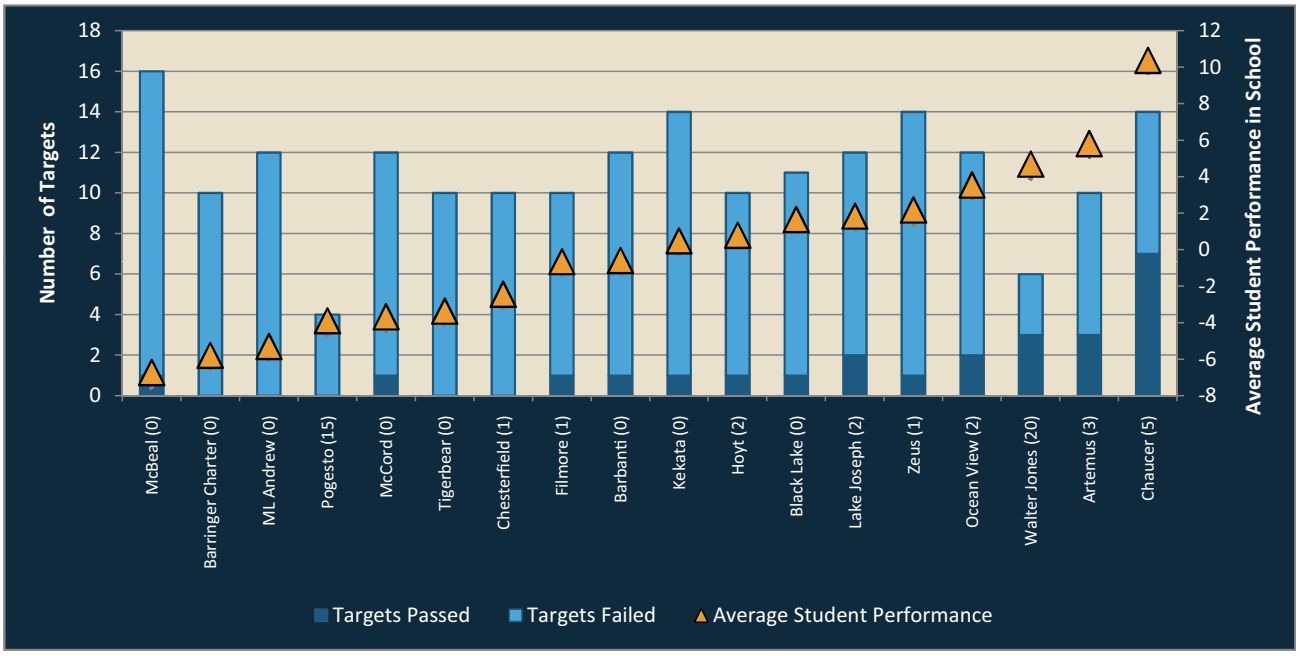

Figure 4. AYP performance of the middle school sample under Massachusetts's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Massachusetts AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Chaucer, for example, met half its targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Massachusetts's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Massachusetts's 2008 AYP rules. **Only one elementary school made AYP while seventeen failed to make it.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The only elementary school (Roosevelt) that made AYP had just one subgroup, which resulted in only four targets for the school to meet (two targets for the overall population in reading and math, and two more targets for the white subgroup in reading and math).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Massachusetts AYP rules. None of the 18 middle schools made AYP.

Where Do Schools Fail?

Figure 3 shows that having few targets is crucial to making AYP, but neither Figures 3 or 4 indicates which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's

Table 2. Elementary school subgroup performance of sample schools under the 2008 Massachusetts AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	46.8%	52.0%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	50.8%	57.9%	N	N			N	N	N	N					N	N			N	N	10	0	0%	N	1
Few	56.3%	60.5%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Nemo	57.4%	70.6%	N	N					N	N									N	N	6	0	0%	N	7
Island Grove	58.5%	71.4%	N	N					N	N					N	N			N	N	8	0	0%	N	5
JFK	64.7%	69.5%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	3
Scholls	70.2%	73.3%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	7
Hissmore	69.7%	74.4%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	7
Wolf Creek	65.6%	73.9%	N	N					N	N					N	N			N	N	8	0	0%	N	5
Alice Mayberry	70.0%	76.9%	N	N					N	N	N	N							Y	Y	8	2	25%	N	9
Wayne Fine Arts	68.2%	85.1%	N	N															N	Y	4	1	25%	N	21
Winchester	70.9%	81.0%	N	N															N	N	4	0	0%	N	22
Coastal	75.3%	77.5%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	3
Paramount	76.4%	79.4%	N	N					N	N					N	N			Y	Y	8	2	25%	N	7
Forest Lake	83.8%	86.2%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	8
Marigold	83.0%	85.6%	Y	Y	N	N			N	N									Y	Y	8	4	50%	N	10
Roosevelt	84.5%	92.1%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	83.9%	90.1%	Y	Y	N	N			N										Y	Y	7	4	57%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Massachusetts rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Four elementary schools (Forest Lake, Marigold, Roosevelt, King Richard) met the reading and the math targets for their overall school population.
- Five middle schools met reading targets for their overall population and only one middle school (Chaucer) met its math target for its overall school population.
- Most of the subgroups in both elementary and middle schools failed to meet their targets.

Table 3. Middle school subgroup performance of sample schools under the 2008 Massachusetts AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	45.2%	71.5%	N	N	N	N	N	N	N	N	N	N			N	N	N	N	N	Y	16	1	6%	N	0
Barringer Charter	48.5%	71.4%	N	N	N	N			N	N	N	N			N	N					10	0	0%	N	0
ML Andrew	44.9%	78.1%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Pogesto	44.0%	82.4%	N	N															N	N	4	0	0%	N	15
McCord Charter	48.2%	80.3%	N	N	N	N			N	N	N	N			N	N			N	Y	12	1	8%	N	0
Tigerbear	55.1%	76.0%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	0
Chesterfield	57.9%	79.1%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	1
Filmore	57.9%	83.3%	N	N	N	N			N	N					N	N			N	Y	10	1	10%	N	1
Barbanti	56.1%	79.3%	N	N	N	N	N	N	N	N					N	N			N	Y	12	1	8%	N	0
Kekata	64.5%	82.3%	N	N	N	N	N	N	N	N	N	N			N	N			N	Y	14	1	7%	N	0
Hoyt	63.9%	83.6%	N	N	N	N			N	N	N	N							N	Y	10	1	10%	N	2
Black Lake	68.4%	84.3%	N	N	N	N			N	N	N	N			N				N	Y	11	1	9%	N	0
Lake Joseph	64.6%	86.6%	N	Y	N	N	N	N	N	N					N	N			N	Y	12	2	17%	N	2
Zeus	68.6%	85.1%	N	N	N	N	N	N	N	N	N	N			N	N			N	Y	14	1	7%	N	1
Ocean View	68.4%	90.7%	N	Y	N	N	N	N	N	N					N	N			N	Y	12	2	17%	N	2
Walter Jones	75.1%	86.1%	N	Y					N	N									Y	Y	6	3	50%	N	20
Artemus	75.0%	88.1%	N	Y	N	N			N	N					N	N			Y	Y	10	3	30%	N	3
Chaucer	79.1%	93.4%	Y	Y	N	N	N	N	N	N			Y	Y	N	Y			Y	Y	14	7	50%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Tables 4 and 5 summarize subgroup performance for sample elementary and middle schools, respectively. In examining these, a few points become clear. First, none of the subgroups did very well with the reading and math tests, most likely because **Massachusetts's proficiency standards are among the highest in the nation, and because unlike most other states, it does not use confidence intervals as a tool to boost its reported proficiency rates**. The only subgroups within the sample elementary and middle schools that ever reached their targets are the white and Asian subgroups (with the exception of one Hispanic subgroup at Chaucer)—neither of which is traditionally academically disadvantaged. It is likely that as NCLB's 100% proficiency deadline approaches, schools in Massachusetts will face increasing sanctions because of their current high standards.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that schools that failed in the majority of other states failed in Massachusetts too.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Massachusetts AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	8	8
Students with limited English proficiency	4	4	4
Low-income students	15	15	14
African-American students	5	5	5
Asian/Pacific Islander students	0	0	0
Hispanic students	7	7	7
American Indian/Alaska Native students	0	0	0
White students	16	9	8

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Massachusetts AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	7	7	7
Low-income students	17	17	17
African-American students	10	10	10
Asian/Pacific Islander students	1	0	0
Hispanic students	13	13	11
American Indian/Alaska Native students	1	1	1
White students	17	14	4

Nevertheless, Massachusetts does produce some anomalies. Winchester and Wayne Fine Arts Elementary Schools both made AYP in the majority of the other states examined, but failed in Massachusetts. The same pattern holds true for Walter Jones Middle School. **These failures are almost certainly the consequence of Massachusetts's higher proficiency standards and lack of confidence intervals, compared to the other states examined. In fact, the only school within our sample that made AYP under the Massachusetts rules was Roosevelt Elementary, which had a much smaller proportion of traditionally academically disadvantaged students (e.g., low income) and far fewer subgroups (and hence, fewer targets to meet) (see Table 6).**

Concluding Observations

This study examined the test performance data of stu-

Table 6. Comparisons between schools that did and didn't make AYP in Massachusetts

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	1	17	0	18
Average student body size	262	307	n/a	859
Average % low income	13	48	n/a	45
Average % nonwhite	19	42	n/a	44
Average performance†	8.85	0.78	n/a	-0.05
Average % growth‡	103	116	n/a	98
Average number of targets to meet	4	8	n/a	11

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

dents from 18 elementary and 18 middle schools across the country to see how these schools would fare under Massachusetts's AYP rules (and AMOs) for 2008. Among this sample, only one elementary school and no middle schools–one from a sample of 36–would have made AYP in Massachusetts. Looking across the 28 state accountability systems examined in the study, this puts Massachusetts at the very low end of the sample distribution in terms of the number of schools making AYP (see Figure 1). Massachusetts' high proficiency standards (and lack of confidence intervals to boost proficiency rates) will mean that schools will have increasing difficulty in meeting the 100% proficiency requirements of NCLB by 2014.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Massachusetts's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income students. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. In the case of Massachusetts, is it "fair" that a state is penalized for having rigorous proficiency standards and annual targets? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Michigan

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Michigan's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Michigan's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Michigan's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Michigan's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **8 of 18 elementary schools** and **14 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Michigan's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, **we find that the number of elementary schools that made AYP in Michigan is exceeded in just 4 other sample states (California, Texas, Arizona, Wisconsin). In addition, Michigan is one of just a handful of states where four or more middle schools made AYP (see Figure 1).**[2]
- Every school in our sample that failed to make AYP in Michigan met expected targets for their overall population but failed because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[3]
- Seven sample schools that made AYP in Michigan failed to make AYP in most other states. This is likely because Michigan's proficiency standards are relatively easy, compared to other states, and these schools generally have fewer accountable subgroups.

Compared with other states in the study, **Michigan** is at the high end of the distribution in terms of how many sample schools make AYP. One could attribute this to a number of factors. First, Michigan's proficiency standards (or cut scores) are relatively easy compared to other states in the study (none are above the 35th percentile according to NWEA norms). An additional factor is that unlike most states, which apply a confidence interval (margin of error) to measurements of group proficiency rates, Michigan applies a standard error to individual student scores. This increases the number of students whose scores are considered passing. A final contributing factor to the large number of schools making AYP in Michigan is that the state applies different annual targets for different grades and subjects (e.g., 54% of grade 8 students in reading are expected to reach proficiency in 2008; that number changes to 65% for grade 3 math students).

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Michigan Educational Assessment Program (MEAP).

[2] It's important to note that Michigan received full and immediate approval from the U.S. Department of Education in 2008 to implement a student growth model in 2007–2008. This analysis, which draws on data from 2005–2006, does not in any way use or incorporate Michigan's student growth model calculations.

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

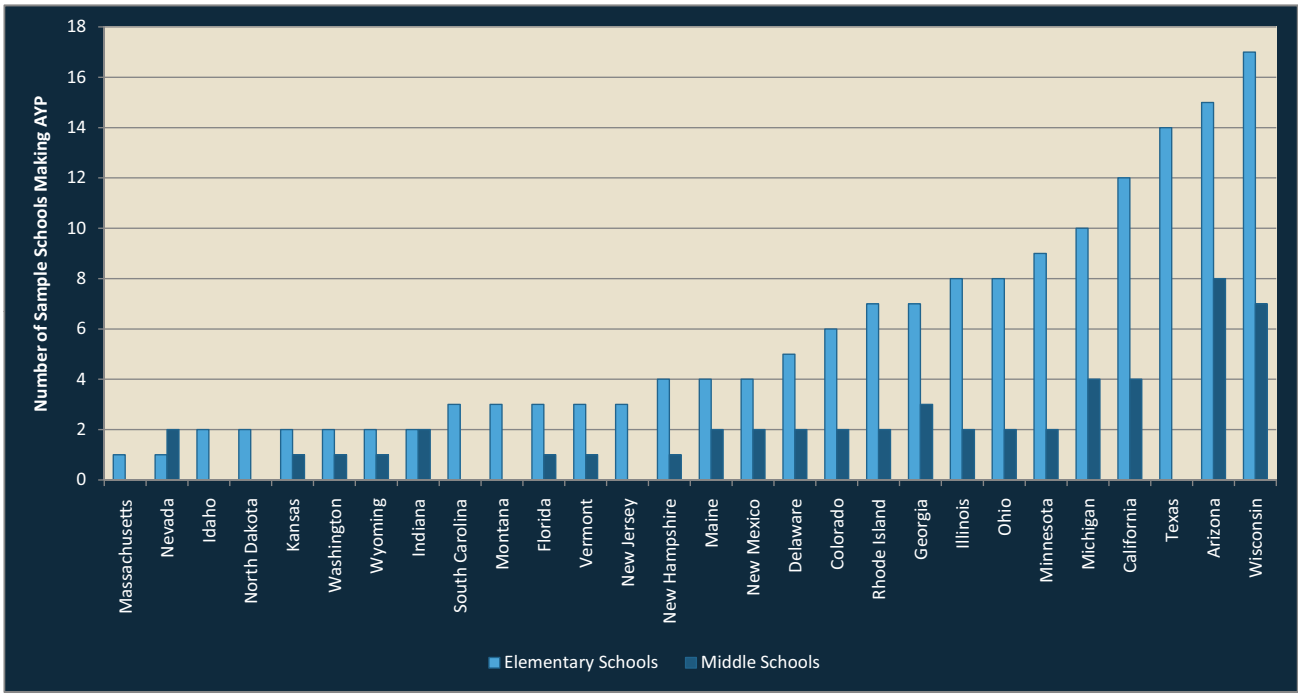

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- Schools with fewer subgroups attained AYP more easily in Michigan than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP. This is the case in other states as well.
- Middle schools had greater difficulty reaching AYP in Michigan than did elementary schools, primarily because their student populations are larger and they therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.
- A strong predictor of a school making AYP under Michigan's system is whether it has enough SWDs to qualify as a separate subgroup. More than half of the schools with enough qualifying SWDs failed to meet their AYP targets.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Michigan's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency stan-

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for limited English proficient (LEP) and SWDs may be slightly more negative than actual findings, mostly because of the differences in testing practices between the Michigan Educational Assessment Program (MEAP), the state assessment, and NWEA's Measures of Academic Progress (MAP), the assessment used in this study. Specifically, the U.S. Department of Education has issued NCLB guidelines permitting schools to exclude small percentages of LEP or disabled students from taking state tests, or providing them alternate assessments. In this study, however, no valid MAP scores were omitted from consideration.

dards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners,[6] among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the Michigan Educational Assessment Program (MEAP) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Michigan's definitions of proficiency ranked below the average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Michigan test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary

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

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

[7] We gave all schools in our sample pseudonyms in this report.

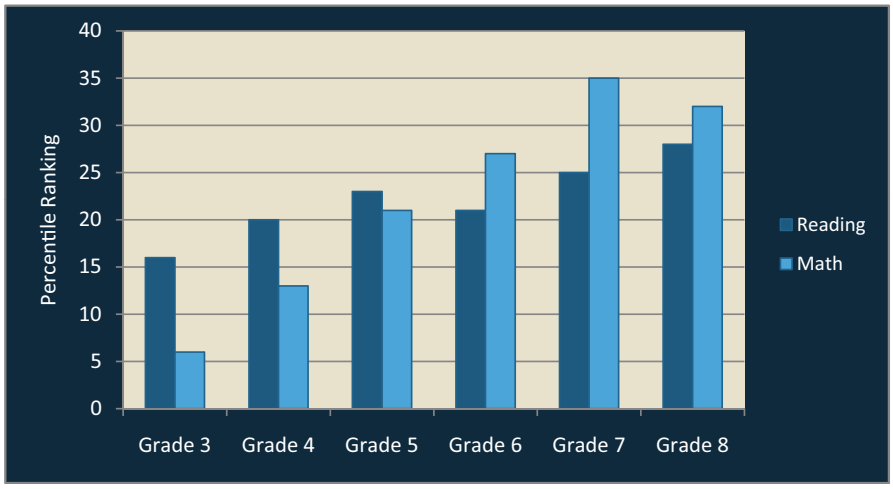

Figure 2. Michigan reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Michigan's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Michigan's cut scores are at or below the 35th percentile.

and 18 middle schools would have fared under Michigan AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Michigan's 2008 AYP rules.

Table 1 shows the pertinent Michigan AYP rules that we applied to elementary and middle schools in this study. Michigan employs a "sliding" minimum subgroup size of 30 or 1% of the school population, whichever is larger, up to a maximum of 200 students.[8] Thirty is a smaller number than is used in most states, which helps ensure that smaller subgroups will still be accountable. Most states, however, employ a fixed number rather than a sliding one, increasing the likelihood that larger schools will be accountable for more subgroups than small schools.

Unlike most states, which apply a confidence interval to measurements of group proficiency rates, Michigan applies standard errors to individual student scores. Technically, this is a more appropriate strategy than using confidence intervals—that is, if the motivation is to correct for test measurement error. However, rather than treating the measurement error correctly (a student's "true" score could be higher OR lower), Michigan merely *adds* the standard error to the student's score, making it easier for students to achieve proficiency on the state test (thus the technical advantage of using standard errors over confidence intervals is lost). Ironically enough, all of the states in the study that use confidence intervals follow essentially this same practice, by treating the margin of error as if it only went in one direction—the one favoring school outcomes. Strictly speaking, such practices cannot be justified purely by a desire to correct for measurement error, because measurement error is seldom unidirectional.

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

[8] In Michigan, the minimum subgroup size is generally 1% of the total school population. Overall, this means that the subgroup size grows with the school size. However, there's also a clause that specifies the minimum subgroup size can't be less than 30 or more than 200. For example, a school with a total population of 3900 would have a minimum subgroup size of 39 (i.e., 1%), but a school with only 900 students would have a minimum subgroup size of 30, since 1% of 900 (i.e., 9) is below the minimum. Similarly, a hypothetical school of 25,000 would have a minimum subgroup size of 200, since 1% of 25,000 (i.e., 250) is greater than the maximum value.

Table 1. Michigan AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 1% of school population, but can't be less than 30 or more than 200	
	SWDs: 1% of school population, but can't be less than 30 or more than 200	
	Low-income students: 1% of school population, but can't be less than 30 or more than 200	
	LEP students: 1% of school population, but can't be less than 30 or more than 200	
CI	**Applied to proficiency rate calculations?**	
	Not used, but 2 standard errors added to individual test scores	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	38	59
Grade 4	38	59
Grade 5	38	59
Grade 6	31	54
Grade 7	31	54
Grade 8	31	54
MATH		
Grade 3	47	65
Grade 4	47	65
Grade 5	47	65
Grade 6	31	54
Grade 7	31	54
Grade 8	31	54

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Michigan's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Michigan's 2008 AYP rules. **Ten elementary schools made AYP and eight failed to make it**. The triangles in the figure show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The majority of the

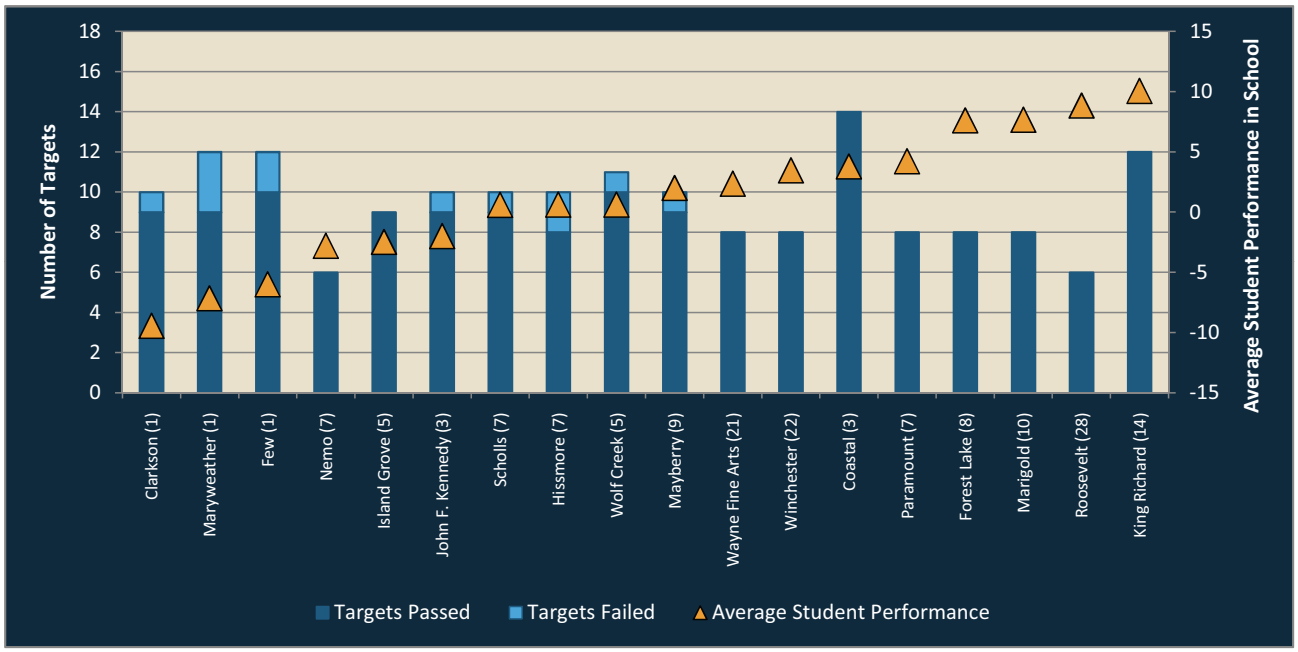

Figure 3. AYP performance of the elementary school sample under the Michigan 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Michigan's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Mayberry Elementary, for example, met 9 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

schools making AYP are in the right half of the figure, meaning that the highest performing students were found at these schools.

Of the schools with lower performing students, the only ones that made AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet. For example, Nemo and Island Grove made AYP but have only six and nine targets each, respectively. Each had to make AYP for its overall student population in reading and math (two targets), for its low-income population (two targets), and for its white population (two more targets). Island Grove also had to make AYP for its LEP population in reading (one target) and for its Hispanic population (two targets).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Michigan AYP rules. **Of 18 in our sample, only 4 made AYP**—one low-performance school (Pogesto), one middle-performance school (Hoyt), and two high-performance schools (Walter Jones and Chaucer). All but Chaucer (the highest performing school in the sample) have relatively few qualifying subgroups.

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or

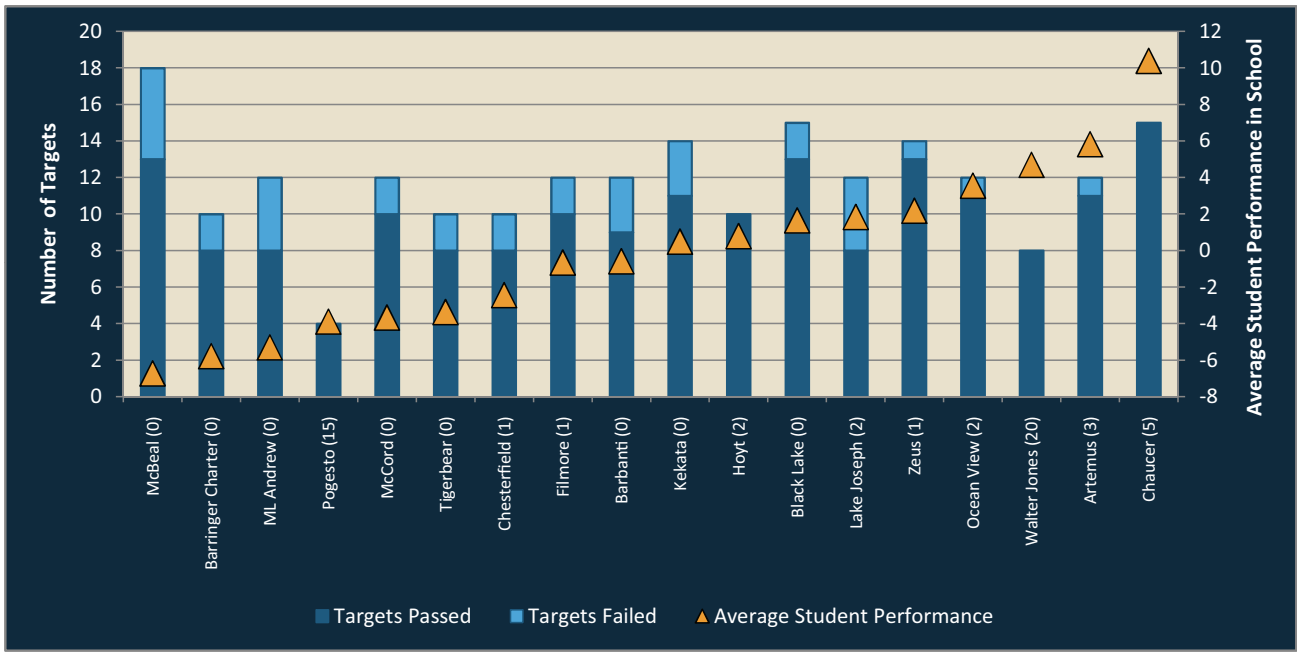

Figure 4. AYP performance of the middle school sample under the Michigan 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Michigan's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Artemus, for example, met 11 of its 12 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP purposes include SWDs, LEP students, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the Michigan rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- All elementary and middle schools met reading and math targets for their overall populations (again, most likely because of Michigan's relatively easy proficiency standards compared to other states).
- Six of the 8 failing elementary schools (Clarkson, JFK, Scholls, Hissmore, Wolf Creek, Alice Mayberry) and 6 of the 14 failing middle schools (Barringer, Tigerbear, Chesterfield, Filmore, Black Lake, and Artemus) missed AYP only for the SWD subgroup.
- Two middle schools (Zeus and Ocean View) fail only because of their LEP subgroups.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. We can see that elementary students did well on Michigan's math test and middle school students performed better in reading than math. This may be because Michigan's proficiency scores are easier in math than in reading at the elementary grades and easier in reading than in math at the middle grades (see Figure 2). Second, the performance of SWDs is proving challenging for schools under Michigan's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* size. Finally, we see that low-income and minority subgroups performed relatively well under Michigan's accountability system.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Michigan AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	88.2%	74.1%	Y	Y	Y	N	Y	Y	Y	Y					Y	Y					10	9	90%	N	1
Maryweather	88.1%	74.4%	Y	Y	N	N	Y	N	Y	Y					Y	Y			Y	Y	12	9	75%	N	1
Few	90.4%	77.7%	Y	Y	Y	N	Y	N	Y	Y					Y	Y			Y	Y	12	10	83%	N	1
Nemo	91.6%	89.8%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	93.7%	87.2%	Y	Y				Y	Y	Y					Y	Y			Y	Y	9	9	100%	Y	4
JFK	96.3%	86.2%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	3
Scholls	96.6%	88.1%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	7
Hissmore	94.3%	90.1%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	92.7%	88.6%	Y	Y	Y	N		Y	Y	Y					Y	Y			Y	Y	11	10	91%	N	5
Alice Mayberry	97.2%	92.4%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	9
Wayne Fine Arts	97.7%	97.7%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	96.7%	94.3%	Y	Y	Y	Y									Y	Y			Y	Y	8	8	100%	Y	22
Coastal	94.5%	88.5%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	3
Paramount	92.9%	89.9%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	7
Forest Lake	98.9%	95.2%	Y	Y	Y	Y			Y	Y									Y	Y	8	8	100%	Y	8
Marigold	99.3%	96.0%	Y	Y	Y	Y			Y	Y									Y	Y	8	8	100%	Y	10
Roosevelt	99.7%	98.6%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	97.6%	97.3%	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y			Y	Y	12	12	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Michigan's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Michigan, too.

But Michigan is also home to a few anomalies. First, consider Island Grove Elementary (see Figure 3). It failed to make AYP in 24 of the 28 states in our sample, yet made AYP in Michigan. In examining Table 2, we can see that Island Grove didn't meet the minimum numbers for the SWD subgroup, which created difficulty for so many other schools within the sample. With fewer accountable subgroups, and with relatively easy proficiency standards (Figure 2), Island Grove made AYP, even when other schools with higher average performance didn't.

Second, look at Pogesto Middle School (see Figure 4). Even with its relatively low average performance, it made

Table 3. Middle school subgroup performance of sample schools under the 2008 Michigan AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	68.8%	73.9%	Y	Y	N	N	N	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	18	13	72%	N	0
Barringer Charter	83.3%	83.9%	Y	Y	N	N			Y	Y	Y	Y			Y	Y					10	8	80%	N	0
ML Andrew	70.6%	82.1%	Y	Y	N	N			N	Y	N	Y			Y	Y			Y	Y	12	8	67%	N	0
Pogesto	70.4%	85.2%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	73.0%	84.7%	Y	Y	N	Y			Y	Y	N	Y			Y	Y			Y	Y	12	10	83%	N	0
Tigerbear	77.8%	80.7%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	0
Chesterfield	82.8%	84.6%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	1
Filmore	82.5%	89.4%	Y	Y	N	N	Y	Y	Y	Y					Y	Y			Y	Y	12	10	83%	N	1
Barbanti	75.7%	82.9%	Y	Y	N	N	N	Y	Y	Y					Y	Y			Y	Y	12	9	75%	N	0
Kekata	84.3%	84.2%	Y	Y	N	Y	N	N	Y	Y	Y	Y			Y	Y			Y	Y	14	11	79%	N	0
Hoyt	87.0%	88.6%	Y	Y	Y	Y			Y	Y	Y	Y							Y	Y	10	10	100%	Y	2
Black Lake	87.7%	87.9%	Y	Y	N	N	Y		Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	15	13	87%	N	0
Lake Joseph	85.2%	89.7%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	88.4%	88.6%	Y	Y	Y	Y	N	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	13	93%	N	1
Ocean View	89.6%	93.7%	Y	Y	Y	Y	N	Y	Y	Y					Y	Y			Y	Y	12	11	92%	N	2
Walter Jones	93.0%	92.6%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	91.5%	90.7%	Y	Y	Y	N			Y	Y			Y	Y	Y	Y			Y	Y	12	11	92%	N	3
Chaucer	93.4%	95.9%	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y			Y	Y	15	15	100%	Y	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

AYP in Michigan, but failed to do so in 13 of 28 states. Like Island Grove, its AYP success in Michigan is likely attributable to the relatively small number of targets (four) it has to meet (shown in Table 3), along with Michigan's relatively easy proficiency standards, compared to other states.

This is consistent with the patterns shown in Table 6, which compares schools that did and didn't made AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower percentages of academically disadvantaged (e.g., low-income) students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Michigan's AYP rules (and AMOs) for 2008. Among this sample, 10 elementary schools and 4 middle schools—

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Michigan AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	13	2	8
Students with limited English proficiency	5	0	2
Low-income students	17	0	0
African-American students	6	0	0
Asian/Pacific Islander students	0	0	0
Hispanic students	9	0	0
American Indian/Alaska Native students	0	0	0
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Michigan AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	11	10
Students with limited English proficiency	9	6	3
Low-income students	17	1	0
African-American students	11	3	0
Asian/Pacific Islander students	4	0	0
Hispanic students	14	0	0
American Indian/Alaska Native students	1	0	0
White students	17	0	0

14 out of a sample of 36—would have made AYP in Michigan. Looking across the 28 state accountability systems examined in the study, this puts Michigan at the high end of the sample distribution in terms of the number of schools making AYP (see Figure 1). In addition, several sample schools made AYP in Michigan that failed to make AYP in most other states, most likely because **Michigan's proficiency standards are relatively easy compared to other states** and its schools generally have fewer accountable subgroups.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Michigan's NCLB accountability system is working exactly as Congress intended: identify-

Table 6. Comparisons between schools that did and didn't make AYP in Michigan, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	10	8	4	14
Average student body size	260	361	586	937
Average % low income	28	69	37	47
Average % nonwhite	29	56	30	48
Average performance†	4.28	-2.59	2.99	-0.93
Average % growth‡	124	104	118	92
Average number of targets to meet	9	11	9	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

ing as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Each of the sample schools made AYP in Michigan for its student populations as a whole. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[9] doesn't the failure of many of these students to meet Michigan's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when sizable numbers of schools (particularly at the middle school level) are unable to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every

[9] See footnote 4.

single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Minnesota

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Minnesota's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Minnesota's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Minnesota's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Minnesota's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **9 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under Minnesota's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, **we find that the number of elementary schools making AYP in Minnesota was exceeded in just 5 other sample states (Michigan, California, Texas, Arizona, and Wisconsin)(see Figure 1).**
- Most of the schools in our sample that failed to make AYP in Minnesota are meeting expected targets for

Compared with other states in the study, **Minnesota** is at the high end of the distribution in terms of how many sample schools make AYP. There are several reasons for this. First, while the majority of states examined apply confidence intervals (margins of error) to their measurements of student proficiency rates, Minnesota uses a "sliding" confidence interval range of 95-99 percent, which is highly unusual. These varying confidence intervals make it easier for Minnesota schools to achieve their targets, with schools that have many subgroups receiving more of a "boost" than schools with fewer targets. Second, Minnesota's minimum subgroup size varies by subgroup. Racial, ethnic, and low-income subgroups have lower minimum *n* sizes than do students with disabilities (SWD) and limited English proficient (LEP) subgroups. Because of this, there are fewer subgroups of SWD and LEP students in elementary schools than in middle schools, which tend to be bigger. Therefore, more elementary schools make AYP. Finally, while most states measure school performance by a proficiency rate (or percentage of students achieving "proficient" or higher on the state test), Minnesota employs a performance "index" which gives partial credit to students attaining "partial proficiency." The resultant score for students in Minnesota is always higher than the actual proficiency percentage (i.e., giving students partial credit for achieving lower proficiency levels is obviously better than no credit, at least for the schools' ratings).

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Minnesota Comprehensive Assessments – Series II.

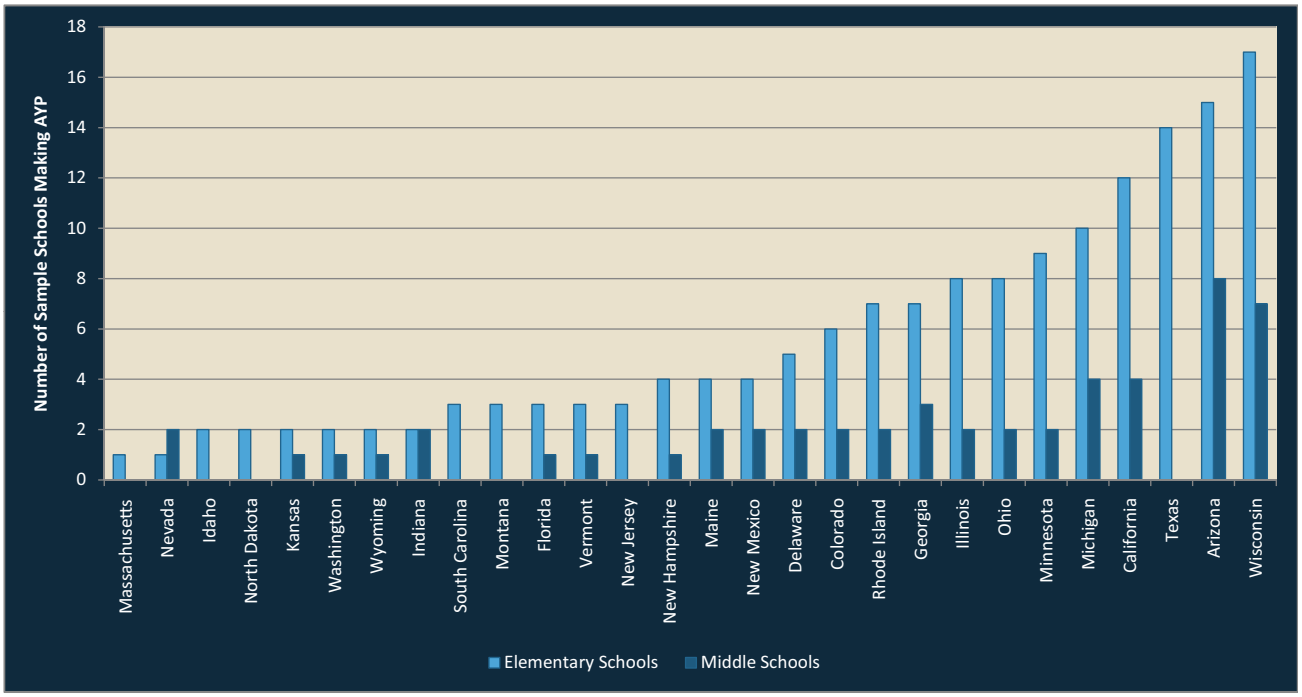

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

their overall populations but failed because of the performance of individual subgroups, particularly English language learners and students with disabilities (SWDs) in middle schools.[2]

- Schools with fewer subgroups attained AYP more easily in Minnesota than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP. This is the case in other states as well.

- Middle schools had greater difficulty reaching AYP in Minnesota than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of whether or not a school would make AYP under Minnesota's system is whether it has enough English language learners to qualify as a separate subgroup. Every school with a limited English proficient (LEP) subgroup failed to make AYP.[3] Likewise, almost all middle schools with enough qualifying SWDs failed to meet their AYP targets.[4]

- **Overall, the application of the confidence interval had some impact on AYP decisions for the sample elementary and middle schools in Minnesota, several of which were assisted in meeting their overall reading and math targets.**

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Minnesota Comprehensive Assessments – Series II, the standardized state test. 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.

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Minnesota's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100. Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discus-

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

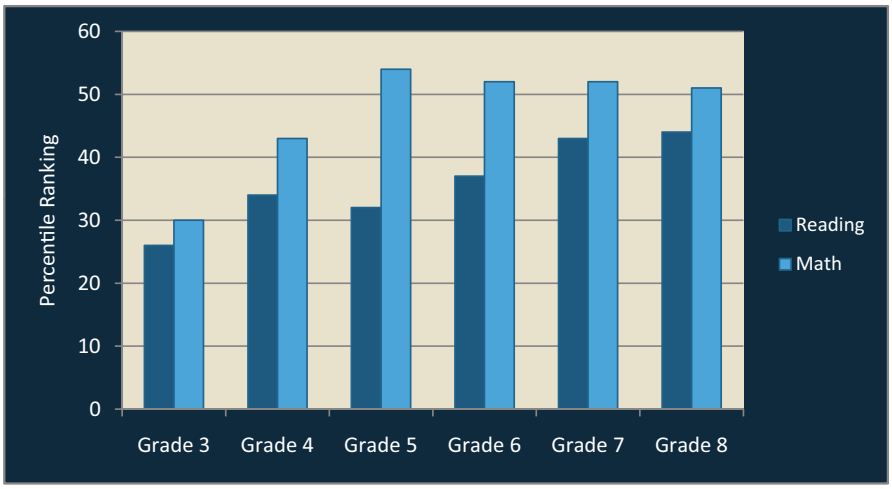

Figure 2. Minnesota reading and math cut score estimates, expressed as percentile ranks (2006))

Note: This figure illustrates the difficulty of Minnesota's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Minnesota's reading cut scores are above the 25th percentile and most of the math cut scores are above the 40th percentile.

Table 1. Minnesota AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 20	
	SWDs: 40	
	Low-income students: 20	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	**Additional notes:**
	Yes, 95%–99% CI, depending on how many subgroups	Confidence interval grows more lenient with more subgroups
AMOs	**Baseline proficiency levels as of 2002 (index)**	**2008 targets (index)**
READING/LANGUAGE ARTS		
Grade 3	72.2	76.2
Grade 4	69.5	73.8
Grade 5	71.9	75.9
Grade 6	70.3	74.5
Grade 7	65.6	70.5
Grade 8	64.0	69.2
MATH		
Grade 3	78.9	81.9
Grade 4	69.6	73.9
Grade 5	59.8	65.5
Grade 6	59.9	65.6
Grade 7	58.8	64.7
Grade 8	58.3	64.3

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

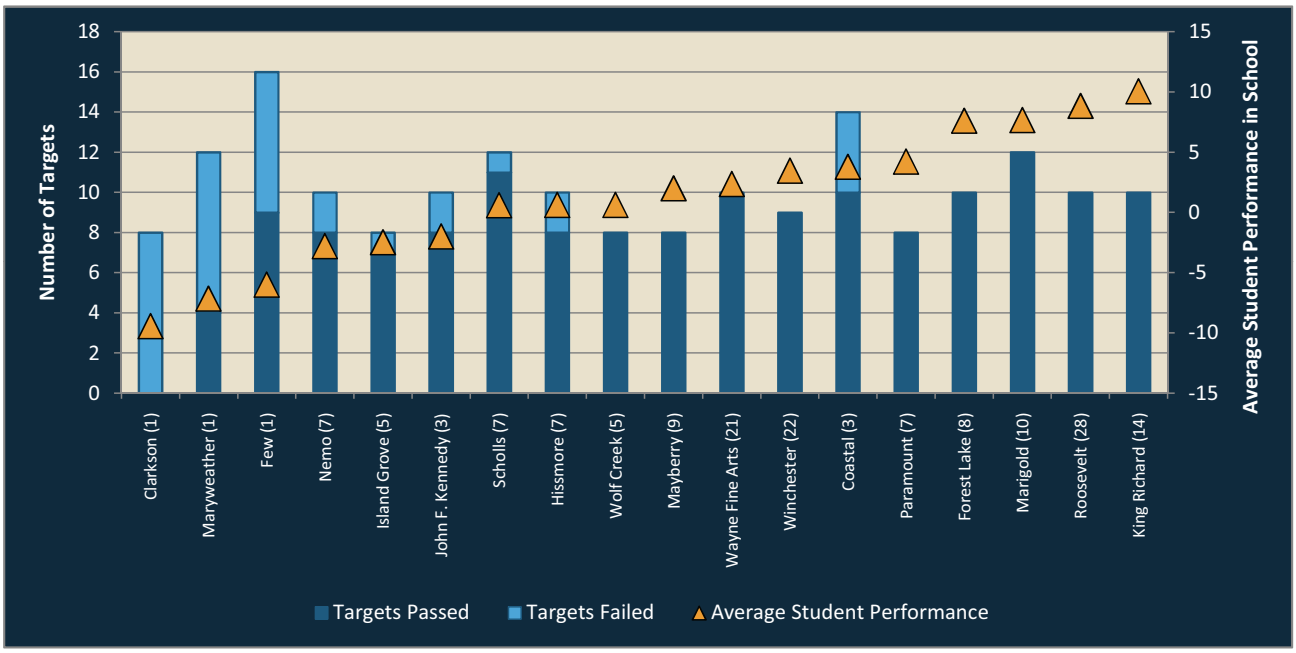

Figure 3. AYP performance of the elementary school sample under Minnesota's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Minnesota's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Hissmore Elementary, for example, met 8 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

sion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Minnesota Comprehensive Assessments – Series II (MCA-IIs) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Minnesota's definitions of proficiency generally ranked above the average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Minnesota test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Minnesota AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Minnesota's 2008 AYP rules.

Table 1 shows the pertinent Minnesota AYP rules that we applied to elementary and middle schools in this study. Minnesota's minimum group size varies by subgroup, with race/ethnic groups and low-income groups at 20 students, and SWDs and LEP groups at 40 students. **Forty is about average, compared to most other states, and 20 is smaller than most.**[7] **This means that schools in Minnesota may have more accountable subgroups than similar schools in other states**. However, because of school size, there are fewer subgroups of SWD and LEP students in elementary schools than in middle schools. This enables more elementary schools to make AYP.

Furthermore, although the majority of states examined in this study apply confidence intervals (margins of sta-

[6] We gave all schools in our sample pseudonyms in this report.

[7] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

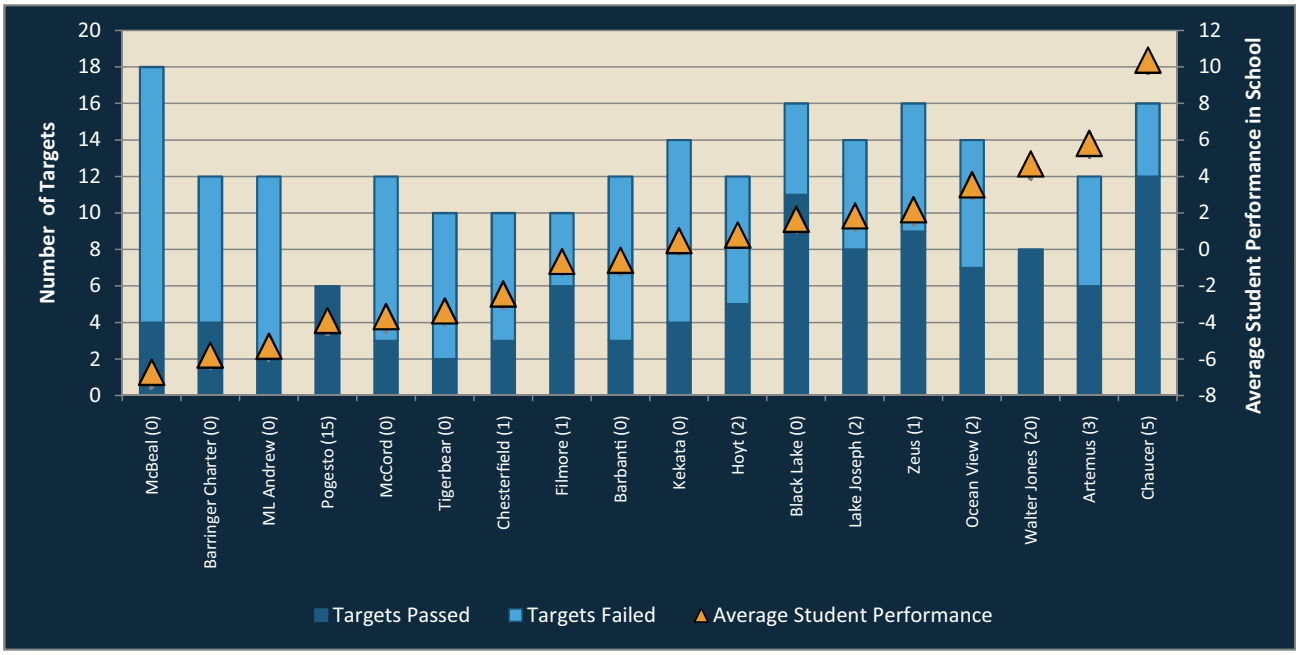

Figure 4. AYP performance of the middle school sample under Minnesota's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Minnesota's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Filmore, for example, met 6 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

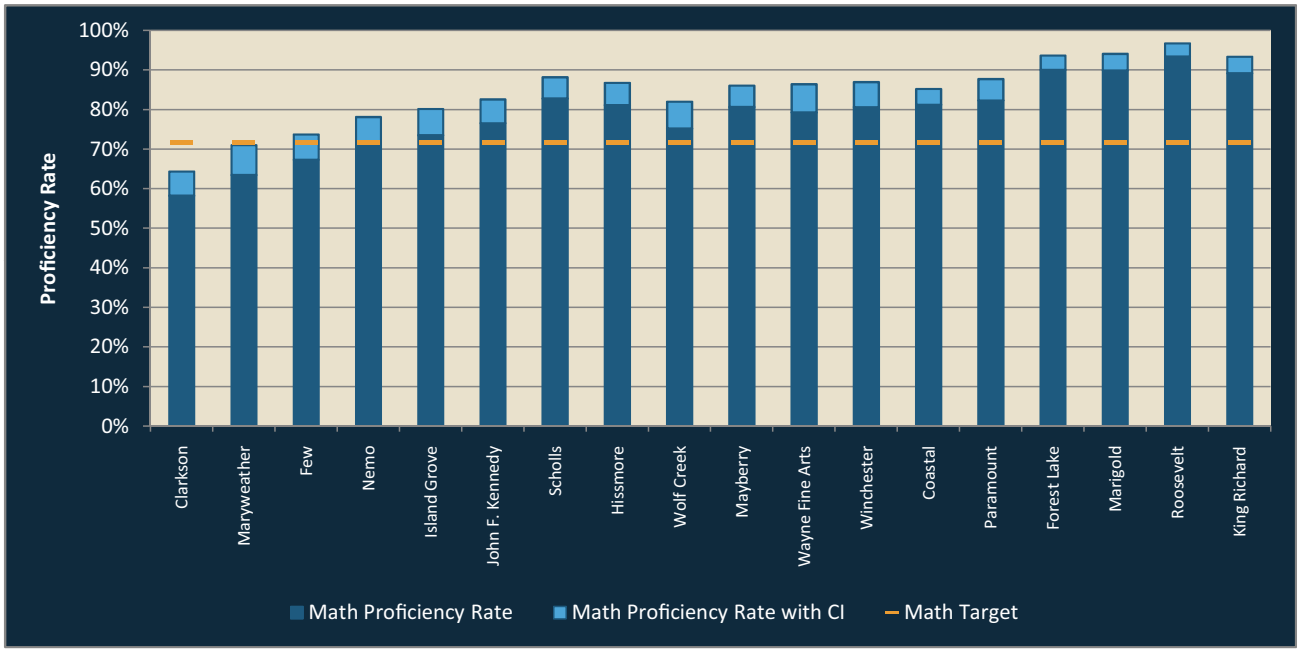

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample elementary schools (Few and Nemo) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

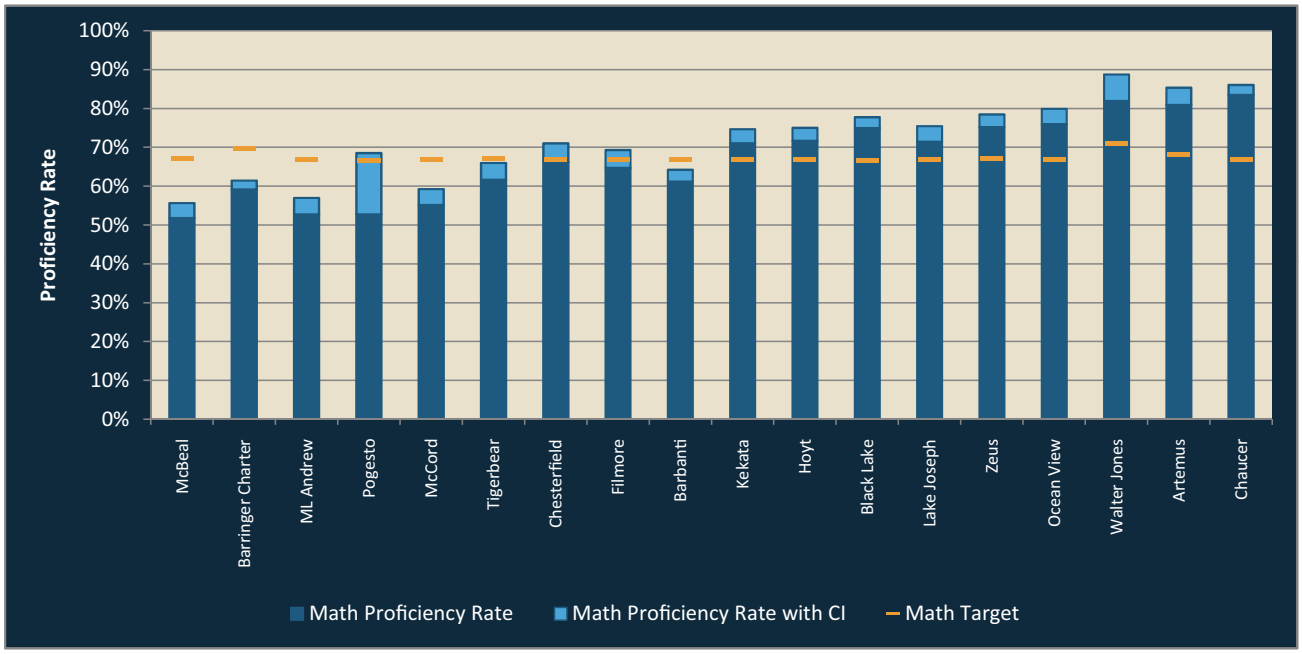

Figure 6. Impact of the confidence interval on middle school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample middle schools (Pogesto and Filmore) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

tistical error) to their measurements of student proficiency rates, **Minnesota's sliding confidence interval range of 95%–99% is unusual. The confidence intervals make it easier for Minnesota schools to achieve their targets, with schools that have many subgroups receiving more of a "boost" than schools with fewer targets.**[8]

Finally, while most states measure school performance by a proficiency rate (or percentage of students achieving a proficient or higher level of performance on the state test), **Minnesota employs a performance index that gives partial credit to students attaining partial proficiency. In the short term, the index makes it easier for schools to meet their targets, although this benefit decreases as the targets approach 100%.**[9]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each

[8] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

[9] Minnesota is one of six states (Massachusetts, New Hampshire, Rhode Island, Vermont, and Wisconsin are the others) in our 28-state sample to use an index that 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.

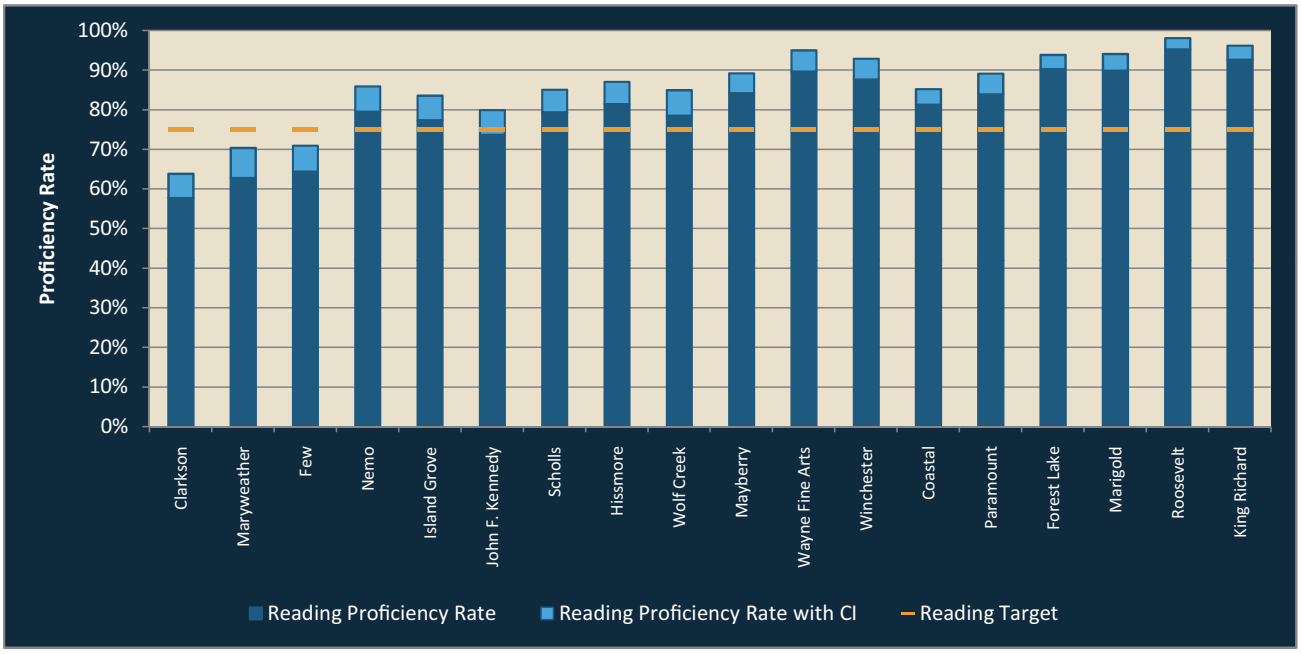

Figure 7. Impact of the confidence interval on elementary school reading proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (JFK) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

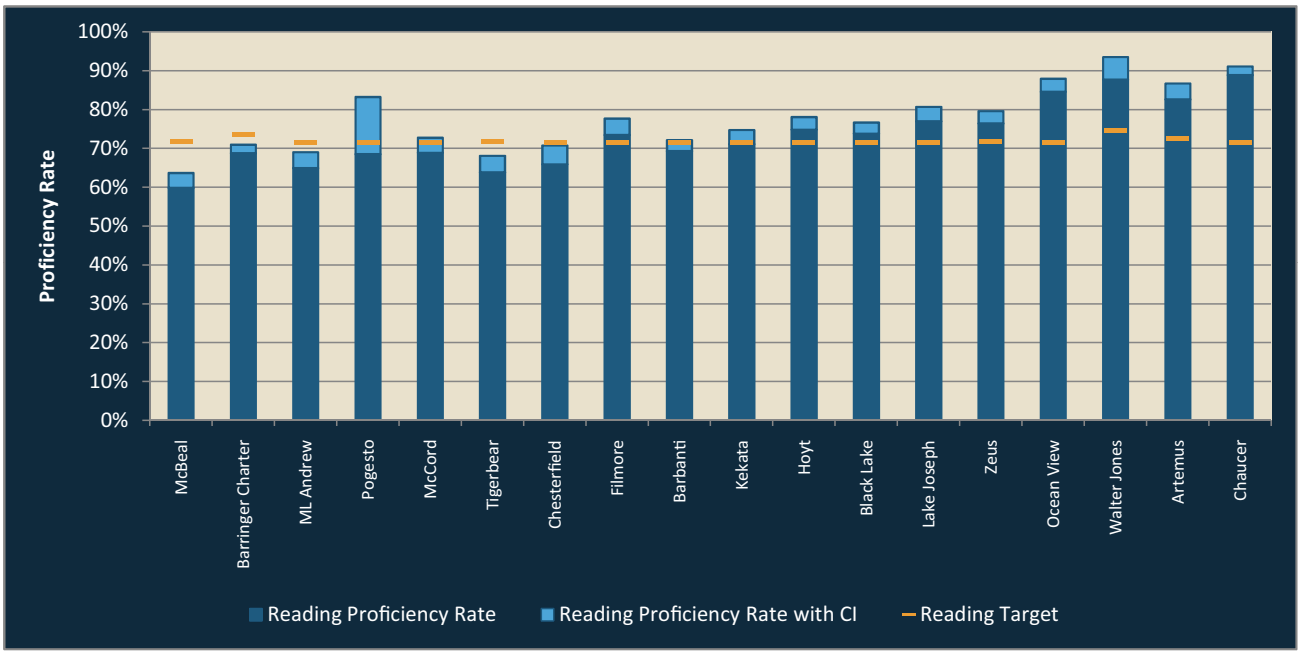

Figure 8. Impact of the confidence interval on middle school reading proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools (Pogesto) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary subgroup performance of sample schools under the 2008 Minnesota AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	58.3%	57.7%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	63.5%	62.8%	N	N			N	N	N	N	Y	Y			N	N			Y	Y	12	4	33%	N	1
Few	67.4%	64.4%	Y	N	N	N	N	N	Y	N	Y	Y			Y	N	Y	Y	Y	Y	16	9	56%	N	1
Nemo	70.9%	79.5%	Y	Y					N	Y	N	Y			Y	Y			Y	Y	10	8	80%	N	7
Island Grove	73.5%	77.4%	Y	Y					Y	Y					Y	N			Y	Y	8	7	88%	N	4
JFK	76.6%	73.7%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	3
Scholls	82.8%	79.4%	Y	Y	Y	N			Y	Y	Y	Y			Y	Y			Y	Y	12	11	92%	N	7
Hissmore	81.2%	81.5%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	75.2%	78.5%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	5
Alice Mayberry	80.7%	84.3%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	79.3%	89.7%	Y	Y					Y	Y	Y	Y			Y	Y			Y	Y	10	10	100%	Y	21
Winchester	80.7%	87.7%	Y	Y					Y	Y				Y	Y	Y			Y	Y	9	9	100%	Y	22
Coastal	81.3%	81.3%	Y	Y	N	N	N	N	Y	Y	Y	Y			Y	Y			Y	Y	14	10	71%	N	3
Paramount	82.3%	84.0%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	7
Forest Lake	90.1%	90.3%	Y	Y	Y	Y			Y	Y	Y	Y							Y	Y	10	10	100%	Y	8
Marigold	89.9%	89.9%	Y	Y	Y	Y			Y	Y			Y	Y	Y	Y			Y	Y	12	12	100%	Y	10
Roosevelt	93.4%	95.3%	Y	Y					Y	Y	Y	Y			Y	Y			Y	Y	10	10	100%	Y	28
King Richard	89.2%	92.7%	Y	Y	Y	Y			Y	Y					Y	Y			Y	Y	10	10	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Minnesota's AYP Rules?

Figure 3 illustrates the AYP performance of the sample **elementary schools** under Minnesota's 2008 AYP rules. **Nine schools out of 18 made AYP**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. Nearly all of the passing schools are in the right half of the figure, meaning that the highest performing students were found at these schools.

Table 3. Middle school subgroup performance of sample schools under the 2008 Minnesota AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	51.8%	59.8%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	59.2%	68.8%	N	N	N	N			N	N	N	N			Y	Y			Y	Y	12	4	33%	N	0
ML Andrew	52.7%	64.9%	N	N	N	N			N	N	N	N			N	N			Y	Y	12	2	17%	N	0
Pogesto	52.8%	68.5%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	15
McCord Charter	55.2%	68.9%	N	Y	N	N			N	N	N	N			N	N			Y	Y	12	3	25%	N	0
Tigerbear	61.7%	63.8%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	0
Chesterfield	66.3%	65.9%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	1
Filmore	64.7%	73.4%	Y	Y	N	N			Y	Y					N	N			Y	Y	10	6	60%	N	1
Barbanti	61.3%	69.3%	N	Y	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	71.1%	71.1%	Y	Y	N	N	N	N	N	N	N	N			N	N			Y	Y	14	4	29%	N	0
Hoyt	71.7%	74.8%	Y	Y	N	N			N	N	N	N			Y	N			Y	Y	12	5	42%	N	2
Black Lake	75.0%	73.8%	Y	Y	N	N			Y	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	16	11	69%	N	0
Lake Joseph	71.5%	77.0%	Y	Y	N	N	N	N	Y	Y	Y	Y			N	N			Y	Y	14	8	57%	N	2
Zeus	75.2%	76.4%	Y	Y	N	N	N	N	Y	N	Y	Y	Y	Y	N	N			Y	Y	16	9	56%	N	1
Ocean View	76.0%	84.6%	Y	Y	N	Y	N	N	N	N			Y	Y	N	N			Y	Y	14	7	50%	N	2
Walter Jones	82.0%	87.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	81.0%	82.6%	Y	Y	N	N			N	N			Y	Y	N	N			Y	Y	12	6	50%	N	3
Chaucer	83.5%	88.9%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Yet almost without regard to average student performance, the schools that made AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet. For example, three out of five schools with the fewest (eight) targets made AYP.

Figure 4 illustrates the AYP performance of the sample **middle** schools under the 2008 Minnesota AYP rules. **Of 18 in our sample, only 2 made AYP**—one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups. Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the darker portions of the bars show the actual proficiency rates at each school, and the lighter portions of the bars show the degree to which these proficiency rates were increased by the application of the confidence interval. The orange lines show the annual measurable objective needed to meet AYP. These figures show that two elementary schools (Few and Nemo) and two middle schools (Pogesto and Filmore) were assisted by the confidence intervals, though all of these except Pogesto still failed to make AYP because of

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Minnesota AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	3	5
Students with limited English proficiency	4	4	4
Low-income students	18	3	3
African-American students	11	1	1
Asian/Pacific Islander students	2	0	0
Hispanic students	14	2	4
American Indian/Alaska Native students	1	0	0
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Minnesota AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	15
Students with limited English proficiency	7	7	7
Low-income students	18	11	13
African-American students	12	9	9
Asian/Pacific Islander students	6	0	0
Hispanic students	15	10	11
American Indian/Alaska Native students	2	1	1
White students	18	0	0

poor subgroup performance (see Figures 3 and 4).

Figures 7 and 8 show the effect of confidence intervals on the reading proficiency rates for elementary and middle schools, respectively. Only one elementary school (JFK) and one middle school (Pogesto) met the overall reading target with the assistance of the confidence interval, but JFK failed to meet all its subgroup targets (see Figure 3). **Overall, the application of the confidence interval provides moderate assistance in helping Minnesota schools achieve their overall math and reading targets.**[10]

[10] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 through 8. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to this report's length and complexity.

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Minnesota rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Only two elementary schools (Clarkson and Maryweather) failed to meet both math and reading targets for their overall school populations, and one additional school (Few) failed to meet its reading target for its overall population.
- Six middle schools failed to achieve their overall math targets and five missed their overall reading targets.
- Two (Scholls and Hissmore) of the nine failing elementary schools missed AYP only because of the SWD subgroup.
- One elementary school (Island Grove) passed in every subgroup except for Hispanic students.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. First, the performance of SWDs is proving to be very challenging for schools under Minnesota's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 40. In fact, nearly all middle schools in the study with qualifying SWD subgroups failed to make AYP. Students with LEP are also struggling to meet the state's targets; every school with a large enough LEP population to qualify as a separate subgroup failed to meet its reading and math targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Minnesota's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Minnesota, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make in Minnesota.

But Minnesota is home to a few anomalies. First, consider Wolf Creek Elementary (see Figure 3). It failed to make AYP in 23 of the 28 states in our sample, yet made AYP in Minnesota. In examining Table 2, we can see that Wolf Creek did not meet the minimum numbers for the LEP or SWD subgroups, which create difficulty for so many other schools within the sample. With fewer accountable subgroups, Wolf Creek made AYP, even when other schools with higher average performance (like Coastal) failed. Second, look at Pogesto Middle School (Figure 4). Even with its relatively low average performance it made AYP in Minnesota, but not in 13 out of 28 states. Like Wolf Creek, its AYP success in Minnesota is most likely attributable to the relatively small number of targets (six) it has to meet.

This is consistent with the patterns shown in Table 6, which compares schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that made AYP did indeed show higher average student performance, but they also

Table 6. Comparisons between schools that did and didn't make AYP in Minnesota, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	9	9	2	16
Average student body size	275	335	124	951
Average % low income	26	67	42	45
Average % nonwhite	30	52	27	46
Average performance†	5.21	-2.76	0.40	-0.11
Average % growth‡	124	106	109	97
Average number of targets to meet	9	11	7	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

differed in the following ways: they had smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower percentages of low income and nonwhite students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Minnesota's AYP rules (and AMOs) for 2008. We found that 9 elementary schools and 2 middle schools–11 in all from a sample of 36—would have made AYP in Minnesota. Looking across the 28 state accountability systems examined in the study, this puts Minnesota at the high end of the sample distribution in terms of the number of schools making AYP (as shown in Figure 1). In addition, **Minnesota's minimum subgroup size varies by particular subgroup, meaning that schools in Minnesota may have more accountable subgroups than similar schools in other states. The application of the confidence interval in Minnesota also provides moderate assistance in helping Minnesota schools achieve their overall math and reading targets**.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Minnesota's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. The majority of the sample schools met the Minnesota math and reading targets for their student populations as a whole, i.e., without considering subgroup performance. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of the student population has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of

making AYP? Doesn't the failure of English language learners, SWDs, low-income students, and other minority groups to meet Minnesota's targets (especially at the middle school level) indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for various subgroups of students, as for other students, but when half or more of schools is not able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Montana

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Montana's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Montana's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Montana's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Montana's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **15 of 18 elementary schools** and **all 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under Montana's accountability system. (This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)

- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools that made AYP in Montana was exceeded in 15 other sample states; Montana ties with 4 other states that each has 3 schools that made AYP (see Figure 1).** Montana also joins Idaho, Massachusetts, South Carolina, and North Dakota with no middle schools that made AYP in the sample.

- Some elementary schools in our sample that failed to make AYP in Montana are meeting expected targets for their overall pupil populations[2] but failed because of the performance of individual subgroups, particularly students with disabilities (SWDs), and English language learners.

- One of the sample middle schools did not make AYP in Montana even though it did so in 23 other states. This may be because some of Montana's annual measurable objectives (AMOs, the proficiency targets needed to make AYP) are relatively high compared to many of the other states examined. In fact, **the way Montana's cut scores and annual targets work together may make it difficult for schools to**

Several factors combine to make **Montana's** AYP rules relatively difficult compared to the other states examined in the study. Montana's proficiency cut scores in math are relatively high, meaning that a student who meets the math proficiency standards in other states might have a harder time doing so in Montana. In addition, the annual targets in Montana are high compared to other states, meaning that schools in Montana must get larger percentages of their students to the "proficient" level than in many other states in order to make AYP. In fact, from our sample of 36 schools, only three elementary and no middle schools met AYP, and none of these three elementary schools had traditionally academically disadvantaged subgroups (such as low income or African American).

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Montana Criterion Referenced Test.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

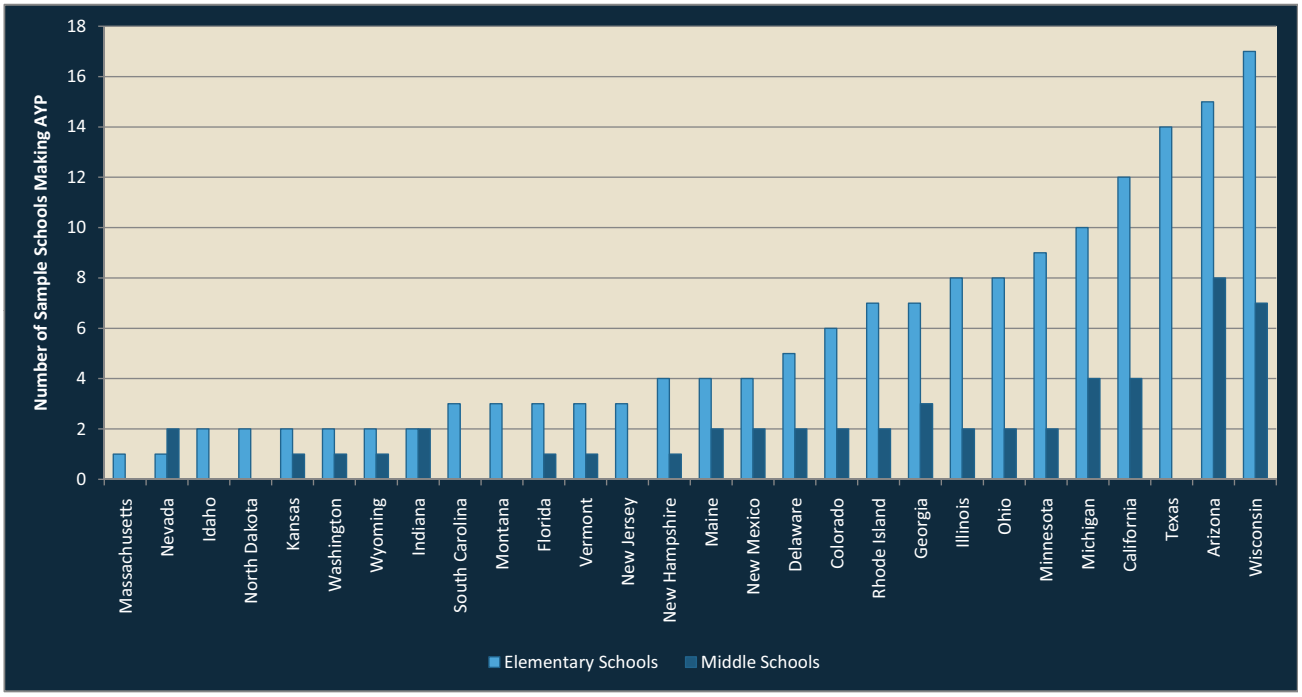

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

make AYP. Specifically, the state's reading cut scores are fairly low but its annual reading targets are demanding; on the other hand, the state's math targets are fairly low, and its math cut scores are somewhat high.

- **In Montana, as in most states, schools with fewer subgroups attained AYP more easily than schools with more subgroups**, even when their average student performance is lower than that in some failing schools. In other words, schools with greater diversity and size face greater challenges in making AYP.
- Montana applies a 95% confidence interval (a statistical margin of error) to its proficiency rate calculations. The confidence interval had little or no impact, however, on final AYP outcomes for sample elementary and middle schools in Montana, partly because sample schools already missed AYP for their subgroup performance.
- As in other states, middle schools in Montana had greater difficulty reaching AYP than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.
- Almost all schools with enough SWDs and limited English proficiency (LEP) students to qualify as separate subgroups failed to meet their targets for those groups.[3]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Montana's tests and those of 25

[3] Note that we use "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup. SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Montana Criterion Referenced Test, the standardized state test. 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.

other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[4] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[5]

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

[5] We gave all schools in our sample pseudonyms in this report.

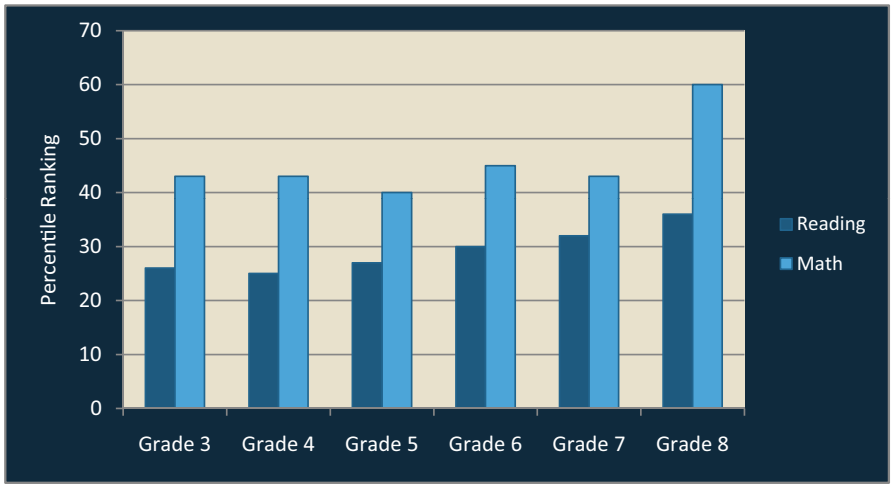

Figure 2. . Montana reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Montana's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Montana's cut scores for reading are below the 40th percentile and all cut scores for math are at or above the 60th percentile.

Table 1. Montana AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 40	
	Low-income students: 40	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	74	83
Grade 4	74	83
Grade 5	74	83
Grade 6	74	83
Grade 7	74	83
Grade 8	74	83
MATH		
Grade 3	51	68
Grade 4	51	68
Grade 5	51	68
Grade 6	51	68
Grade 7	51	68
Grade 8	51	68

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).
Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

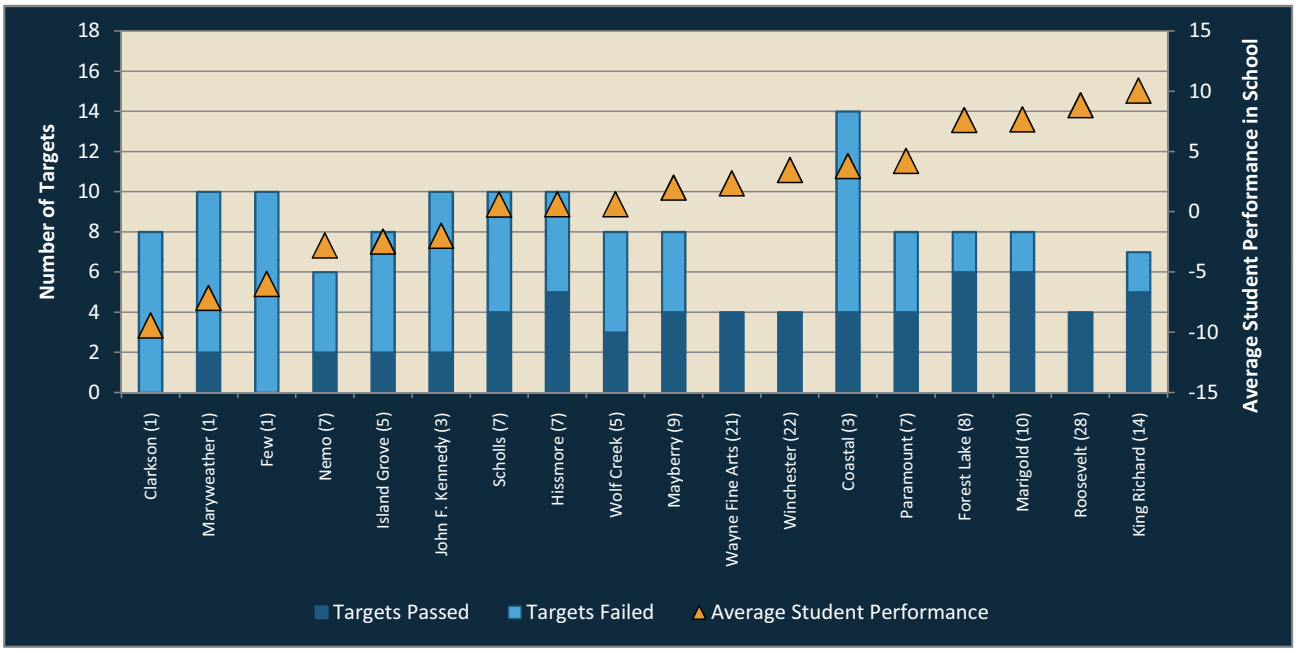

Figure 3. AYP performance of the elementary school sample under Montana's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Montana's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Marigold Elementary, for example, met six of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average

Proficiency cut score estimates for the Montana Criterion-Referenced Test (Montana CRT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Montana's proficiency standards in reading ranked about average compared with the standards set by the other 25 states in that study, and its proficiency standards in math ranked above average. These cut scores were used to estimate whether students would have scored as proficient or better on the Montana test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Montana AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Montana's 2008 AYP rules.

Table 1 shows the pertinent Montana AYP rules that were applied to elementary and middle schools in the current study. Montana's minimum subgroup size is 40, which is about average, compared to most other states we examined.[6]

Furthermore, Montana, like most states, applies a 95% confidence interval (or margin of statistical error) to its measurements of student proficiency rates.[7] So, for instance, even though schools are supposed to get 68% of their grade 3 students to the proficient level on the state math test, as well as 68% of the grade 3 students in each subgroup, applying the confidence interval means that the real target can be lower, particularly with smaller groups.

[6] It's worth noting, however, that schools in Montana are likely to be small and an *n* size of 40, though average, may in fact exclude more subgroups than would be the case in states with larger schools overall.

[7] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

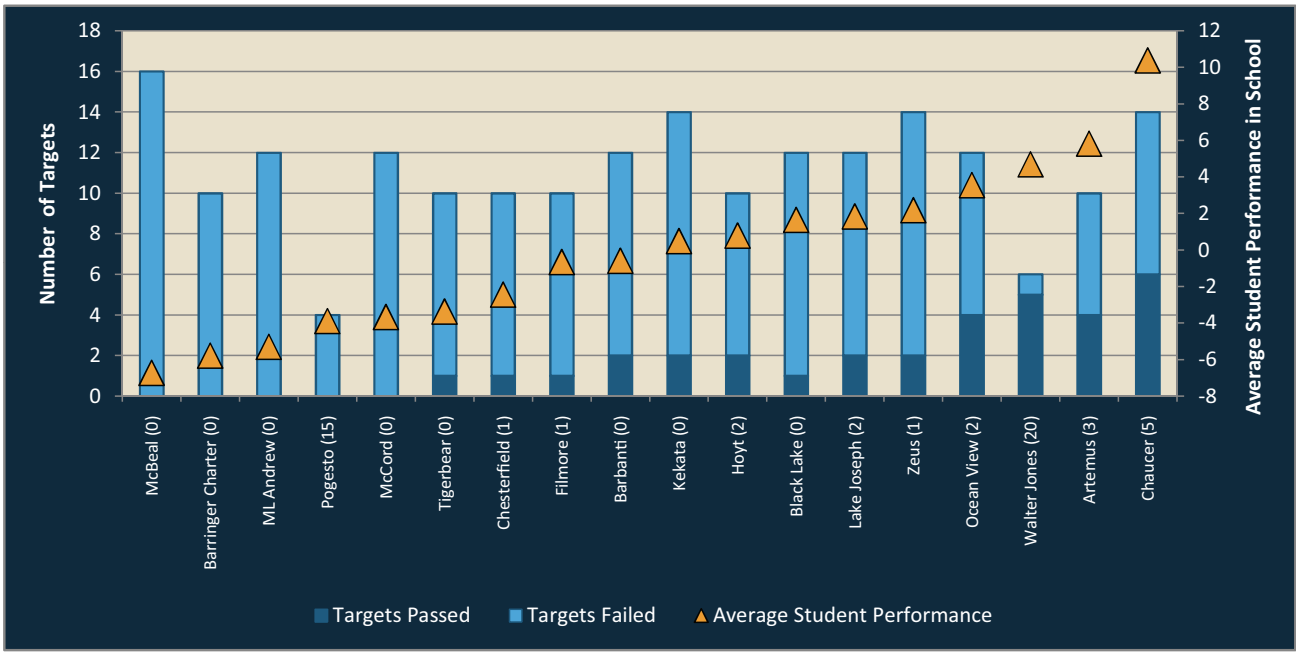

Figure 4. AYP performance of the middle school sample under Montana's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Montana's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Walter Jones Middle School, for example, met five of its six targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

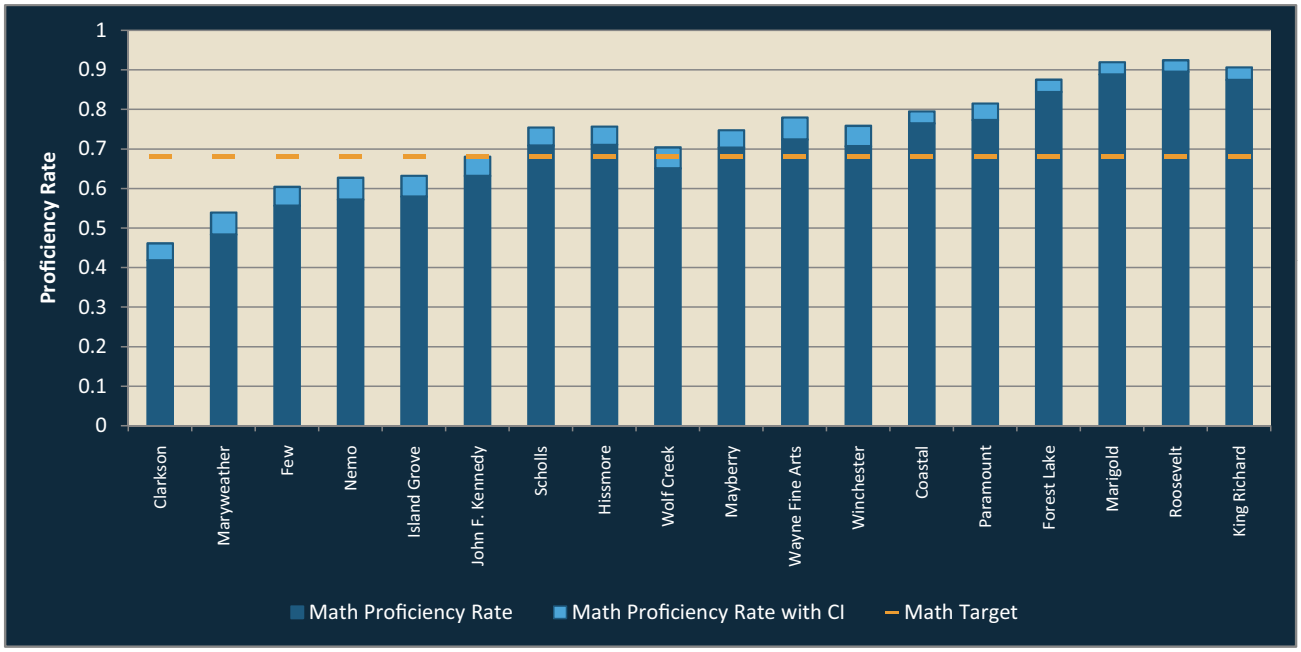

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under Montana's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools, Wolf Creek, was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

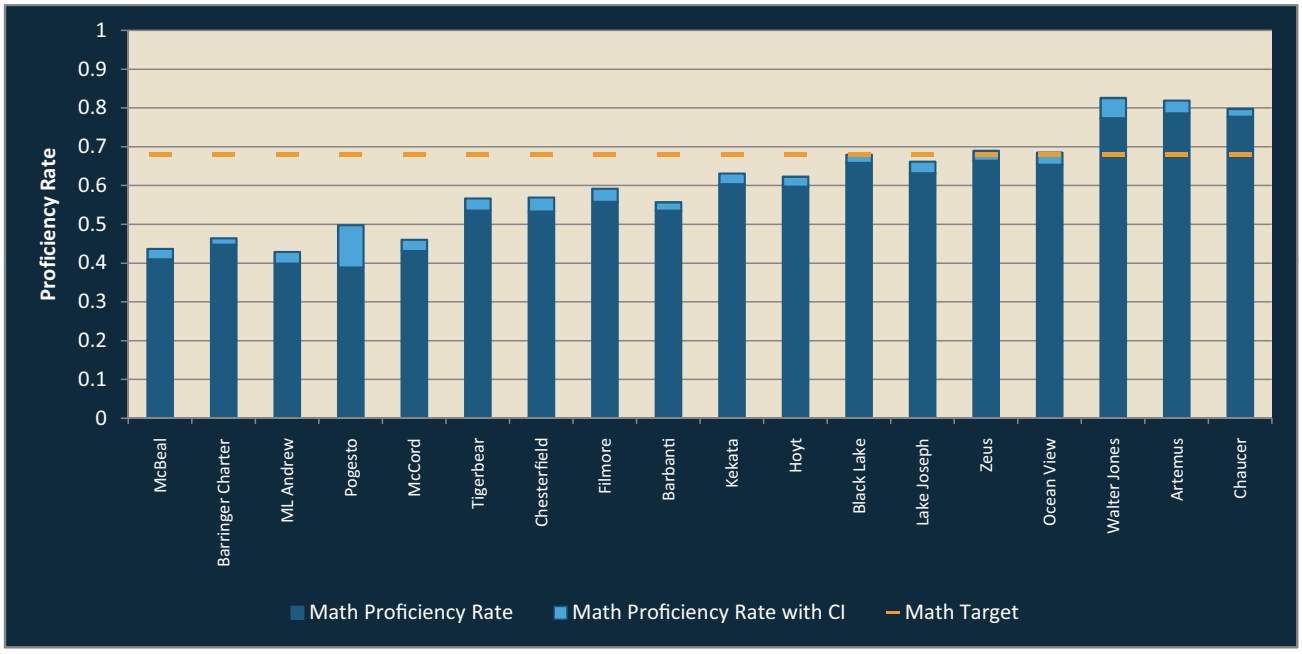

Figure 6. Impact of the confidence interval on middle school math proficiency rates under Montana's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample elementary schools, Black Lake and Zeus, were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Montana's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Montana's 2008 AYP rules. Only 3 elementary schools made AYP while 15 failed to make it. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. All passing schools are in the right half of the figure, meaning that the higher performing students were found at these schools.

Yet almost without regard to average student performance, the only schools made AYP were those with relatively few

Table 2. Elementary school subgroup performance of sample schools under the 2008 Montana AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	41.9%	47.3%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	48.4%	57.1%	N	N			N	N	N	N					N	N			Y	Y	10	2	20%	N	1
Few	55.7%	59.5%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Nemo	57.2%	75.3%	N	N					N	N									Y	Y	6	2	33%	N	7
Island Grove	58.0%	72.4%	N	N					N	N					N	N			Y	Y	8	2	25%	N	4
JFK	63.2%	67.5%	Y	N	N	N			N	N	N	N							Y	N	10	2	20%	N	3
Scholls	70.9%	74.7%	Y	N	N	N			Y	N	N	N							Y	Y	10	4	40%	N	7
Hissmore	71.1%	77.5%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	7
Wolf Creek	65.1%	73.5%	Y	N					N	N					N	N			Y	Y	8	3	38%	N	5
Alice Mayberry	70.3%	80.3%	Y	Y					N	N	N	N							Y	Y	8	4	50%	N	9
Wayne Fine Arts	72.4%	86.8%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	70.8%	83.9%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	76.5%	79.6%	Y	N	N	N	N	N	Y	N	N	N			N	N			Y	Y	14	4	29%	N	3
Paramount	77.3%	79.9%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	7
Forest Lake	84.5%	87.6%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	88.8%	89.5%	Y	Y	Y	N			Y	N									Y	Y	8	6	75%	N	10
Roosevelt	89.6%	94.2%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	87.5%	89.8%	Y	Y	N	N			Y										Y	Y	7	5	71%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

qualifying subgroups—and thus the fewest targets to meet. For example, Wayne Fine Arts and Winchester passed, but had only four targets each. Each must make AYP for its overall student population in reading and math (two targets) and for its white population (two more targets).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Montana AYP rules. **Of 18 middle schools in our sample, none passed**.

Figures 5 and 6 indicate the degree to which schools' overall math proficiency rates are aided by Montana's confidence interval for elementary and middle schools, respectively. On these figures, the darker portion of the bars show the actual proficiency rates at each school, and the lighter portion of the bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the annual measurable objective needed to meet AYP.

These figures show that two elementary schools (JFK and Wolf Creek) and three middle schools (Black Lake, Zeus, and Ocean View) are assisted by the confidence intervals to meet their overall math targets (note how the

Table 3. Middle school subgroup performance of sample schools under the 2008 Montana AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	41.0%	56.8%	N	N	N	N	N	N	N	N	N	N			N	N	N	N	N	N	16	0	0%	N	0
Barringer Charter	44.8%	62.9%	N	N	N	N			N	N	N	N			N	N					10	0	0%	N	0
ML Andrew	39.9%	59.9%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Pogesto	38.9%	68.5%	N	N															N	N	4	0	0%	N	15
McCord Charter	43.2%	63.4%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Tigerbear	53.6%	59.9%	N	N	N	N			N	N	N	N							Y	N	10	1	10%	N	0
Chesterfield	53.3%	63.2%	N	N	N	N			N	N	N	N							Y	N	10	1	10%	N	1
Filmore	55.8%	71.4%	N	N	N	N			N	N					N	N			Y	N	10	1	10%	N	1
Barbanti	53.6%	66.0%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Kekata	60.4%	68.5%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	0
Hoyt	59.7%	72.3%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	2
Black Lake	65.8%	73.3%	N	N	N	N			N	N	N	N			N	N			Y	N	12	1	8%	N	0
Lake Joseph	63.2%	76.3%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	2
Zeus	66.4%	74.3%	Y	N	N	N	N	N	N	N	N	N			N	N			Y	N	14	2	14%	N	1
Ocean View	65.4%	83.7%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	77.3%	85.1%	Y	Y					Y	N									Y	Y	6	5	83%	N	20
Artemus	78.6%	82.0%	Y	Y	N	N			N	N					N	N			Y	Y	10	4	40%	N	3
Chaucer	77.7%	87.9%	Y	Y	N	N	N	N	N	N			Y	Y	N	N			Y	Y	14	6	43%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

orange line falls within the light blue band). Figures 3 and 4 show, however, that all five of these schools still fail to meet some of their subgroup targets. The same is true for reading (not shown). So, although a few schools met their overall targets with the help of the confidence interval, they still missed subgroup targets, and therefore, failed to make AYP. **Overall, the confidence interval had little or no impact on final AYP outcomes for sample elementary and middle schools in Montana.**[8]

Where Do Schools Fail?

Figures 3 and 4 illustrate that a few elementary schools with only middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

[8] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Montana AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	7	8
Students with limited English proficiency	4	4	4
Low-income students	15	9	13
African-American students	5	4	5
Asian/Pacific Islander students	0	0	0
Hispanic students	7	7	7
American Indian/Alaska Native students	0	0	0
White students	16	0	1

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Montana AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	7	7	7
Low-income students	17	16	17
African-American students	10	10	10
Asian/Pacific Islander students	1	0	0
Hispanic students	13	13	13
American Indian/Alaska Native students	1	1	1
White students	17	4	9

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Montana rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Five elementary schools failed to meet both the math and reading targets for their overall school population. Five more elementary schools failed to meet their overall targets in reading.
- Most middle schools failed to meet their overall reading and math targets.
- Two (Forest Lake and King Richard) of the 15 failing elementary schools missed only for the SWD subgroup.
- One middle school (Walter Jones) failed only for its low-income subgroup.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively.[9] First, almost every school with a large enough academically disadvantaged population to qualify as a separate subgroup (e.g., low income, African American, Hispanic) failed to meet its targets for these students. Students with disabilities and limited English proficiency did just as poorly, failing in every elementary or middle school in which that subgroup was accountable. Second, elementary schools did slightly better than middle schools because they have fewer subgroups.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Montana's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Montana, too (though they are the only 3 to do so). Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed in Montana.

But Montana is also home to at least one anomaly. Consider Walter Jones (see Figure 4). It made AYP in 20 of the 28 states in our sample, but not in Montana. In examining Table 3, we can see that Walter Jones failed to meet the reading target for its low-income subgroup. Although Montana's reading cut scores at the middle school grades are fairly low (except at eighth grade), its annual targets are relatively high (i.e., 83% are expected to reach proficiency) compared with many other states. This may account for the fact that this group missed its target, even though it passed in most other states.

Other state reports contain a section comparing some of the characteristics of the sample schools that made AYP versus those that did not. In Montana, none of the sample middle schools made AYP, and among elementary schools, the only striking difference between schools that made AYP and those that didn't is that the former had fewer subgroups.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Montana's AYP rules (and AMOs) for 2008. We found that only 3 elementary schools and no middle schools—3 in all, from of a sample of 36—would have made AYP in Montana. Looking across the 28 state accountability systems examined in the study, this puts Montana in the lower middle part of the sample distribution, as shown in Figure 1. It's worth noting that the way Montana's cut scores and annual targets work together may make it difficult for schools to make AYP.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Montana's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Many of the sample elementary and middle

[9] Recall that elementary schools did better on Montana's math test than middle school students did, perhaps because Montana's proficiency scores are lower in reading (see Figure 2).

schools met their reading and math targets for their student populations as a whole, that is, without considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their students aren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and students with disabilities are more generous under the current state assessment system,[10] doesn't the massive failure of these students to meet Montana's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other pupils, but when almost no school is able to meet the goal perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[10] See footnote 3.

Nevada

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Nevada's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Nevada's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Nevada's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Nevada's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **17 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under Nevada's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students. It's also partly because Nevada's minimum subgroup size is relatively small (25) compared to other states; this means more subgroups are held accountable for performance. In fact, a few sample schools that made AYP in most other states did not make it in Nevada, largely owing to the state's *n* size. (This occurred despite Nevada's fairly low annual performance targets, which require barely half of students to be proficient in math and reading in 2008).
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools that made AYP in Nevada was exceeded by virtually all of the other sample states (Massachusetts and Nevada tie with a single elementary school making AYP). Nevada is one of 10 states with 2 middle schools that made AYP in the sample** (see Figure 1).
- Many schools in our sample that failed to make AYP in Nevada met expected targets for their overall populations[2] but failed because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.
- In Nevada, schools with fewer subgroups attained AYP more easily than schools with more subgroups, even when their average student performance is

A couple of key factors combine to place **Nevada** at the low end of the state distribution in terms of the number of schools making AYP. First, Nevada's definitions of proficiency generally ranked at or above average compared to the standards set by the other 27 states in the study. This means that students had to perform at a higher level in order to be deemed proficient in Nevada. Second, Nevada's minimum subgroup size is relatively small (25), meaning that more subgroups are held separately accountable in Nevada than would be in other states. In fact, every single school with a limited English proficient (LEP) or students-with-disabilities (SWD) subgroup failed to make AYP in Nevada.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Nevada Criterion Referenced Test.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup.

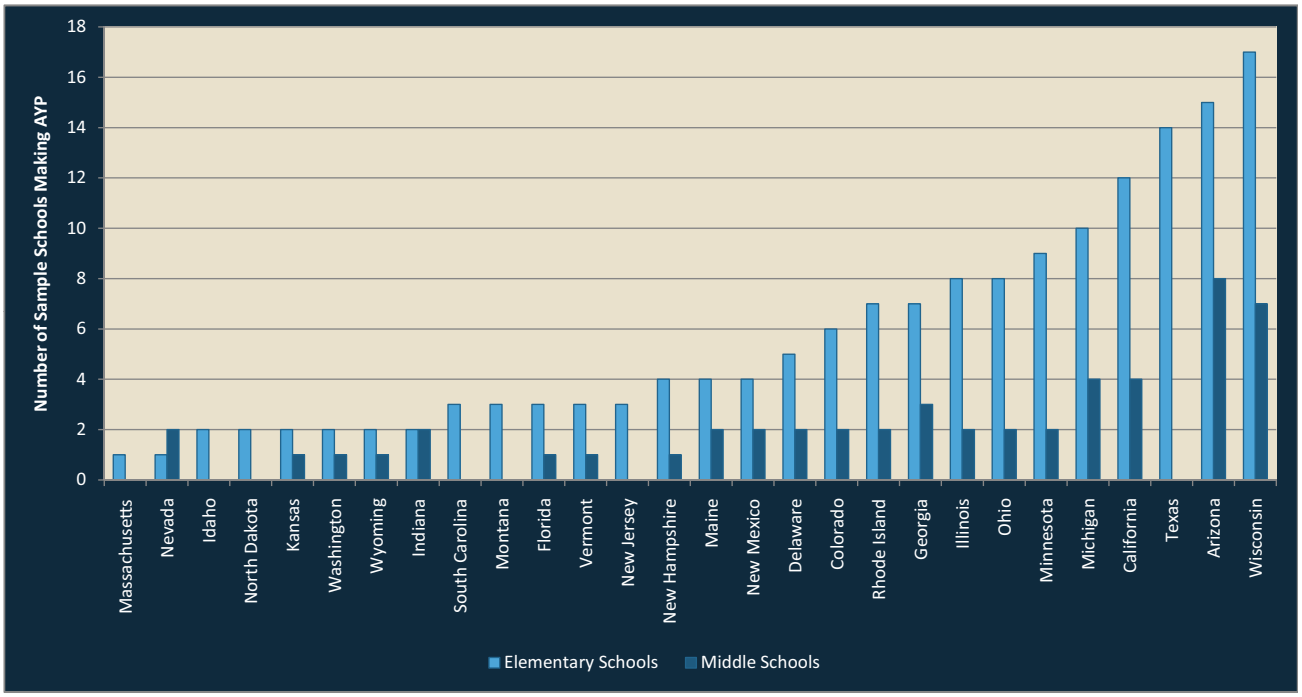

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- Every school with a limited English proficient (LEP)[3] subgroup failed to make AYP. Likewise, almost all schools with enough qualifying SWDs failed to meet their AYP targets.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Nevada's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Nevada Criterion Referenced Test, the standardized state test. 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.

they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than 10 pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Nevada Criterion Referenced Test (Nevada CRT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Nevada's definitions of proficiency generally ranked about average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Nevada test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Nevada AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Nevada's 2008 AYP rules.

Table 1 shows the pertinent Nevada AYP rules that we applied to elementary and middle schools in this study. **Nevada's minimum subgroup size is 25, which is small compared to most other states examined in the study, meaning that Nevada schools will have more accountable subgroups than would similar schools in other states.**[7]

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

[6] We gave all schools in our sample pseudonyms in this report.

[7] It's also possible that Nevada's schools are small and that an *n* size of 25 makes sense for that state.

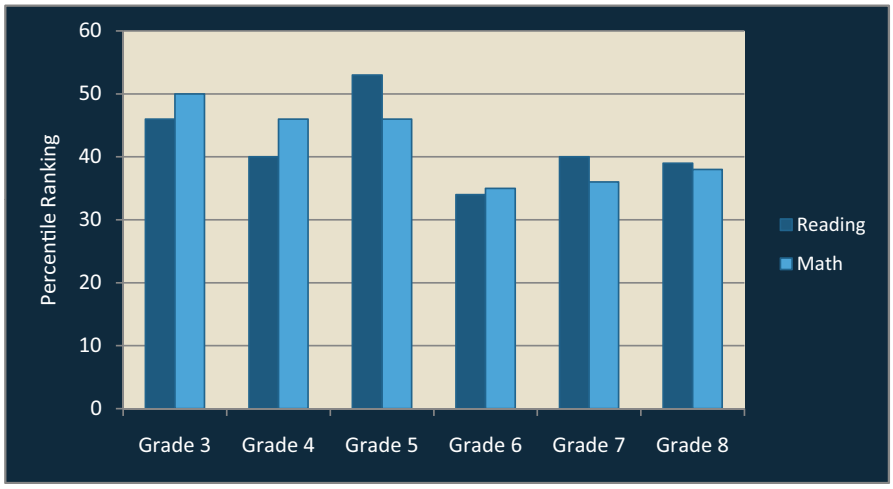

Figure 2. Nevada reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Nevada's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. Almost all of Nevada's cut scores are at or below the 50th percentile.

Table 1. Nevada AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 25	
	SWDs: 25	
	Low-income students: 25	
	LEP students: 25	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	32.4	51.7
Grade 4	32.4	51.7
Grade 5	32.4	51.7
Grade 6	n/a	58.0
Grade 7	n/a	58.0
Grade 8	n/a	58.0
MATH		
Grade 3	37.3	56.3
Grade 4	37.3	56.3
Grade 5	37.3	56.3
Grade 6	n/a	54.6
Grade 7	n/a	54.6
Grade 8	n/a	54.6

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not available

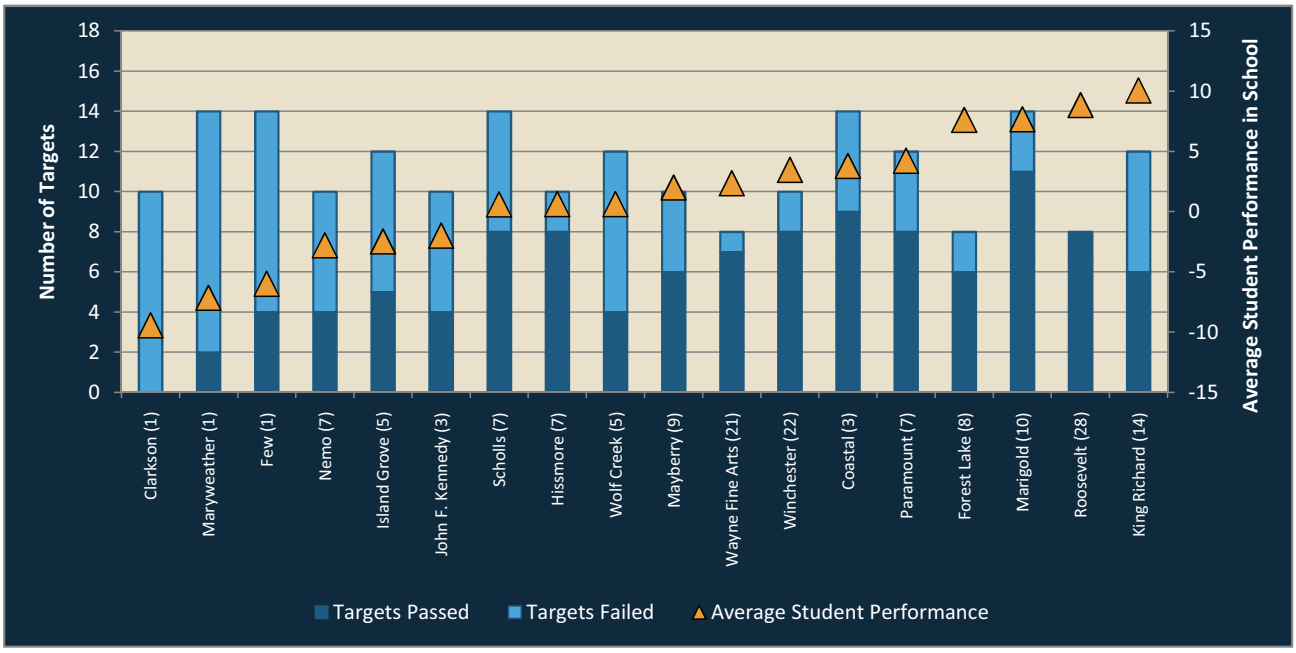

Figure 3. AYP performance of the elementary school sample under Nevada's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Nevada's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Wayne Fine Arts Elementary, for example, met seven of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

Nevada, like the majority of states in the study, applies 95% confidence intervals to its measurements of student proficiency rates.[8] So, for instance, even though schools are supposed to get 51.7% of their grade 3 students to the proficient level on the state reading test (and 51.7% of the grade 3 students in each subgroup), applying the confidence interval means that the real target can be lower, particularly with smaller groups.

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population

[8] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

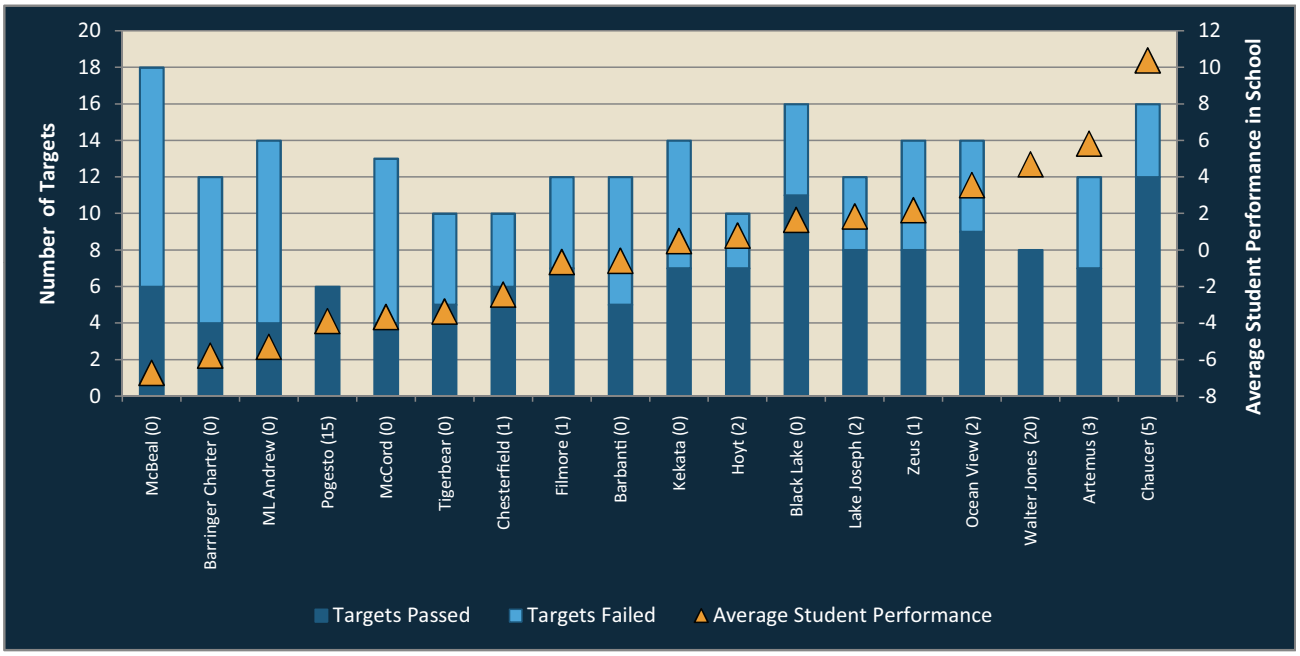

Figure 4. AYP performance of the middle school sample under Nevada's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Nevada's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Hoyt Middle School, for example, met 7 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Nevada's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Nevada's 2008 AYP rules. **Only one elementary school made AYP and 17 failed.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The only school making AYP (Roosevelt) is in the right half of the figure, meaning that relatively high performing students were found at that school. Roosevelt was also one of the only high performing schools with relatively few subgroups (and hence, targets to meet).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Nevada AYP rules. **Of 18 middle schools in our sample, only 2 passed**—one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by Nevada's confidence interval for elementary and middle schools, respectively. On this figure, the darker portions of the bars show the actual proficiency rates at each school and the lighter portions of the bars show the degree to which these proficiency rates were increased by applying the confidence interval. The orange lines show the AMOs needed to meet AYP. These figures show that three sample elementary schools (Few, Nemo, and Island Grove) and two middle schools (ML Andrew and Pogesto) were assisted by the confidence interval. However, all of these schools but Pogesto already failed to make AYP because of low subgroup performance (see Figures 3 and 4), and therefore did not make AYP.

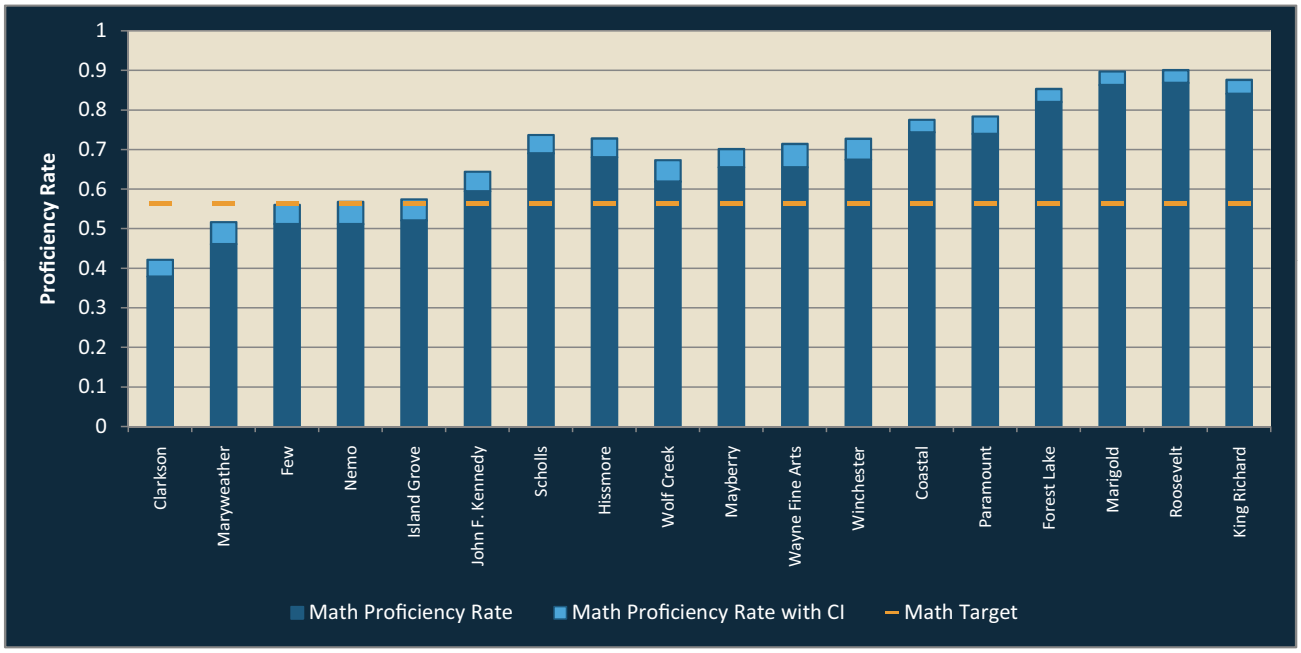

Figure 5. Impact of the confidence interval on elementary school math proficiency rates for 2008

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that three of the sample elementary schools (Few, Nemo, and Island Grove) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

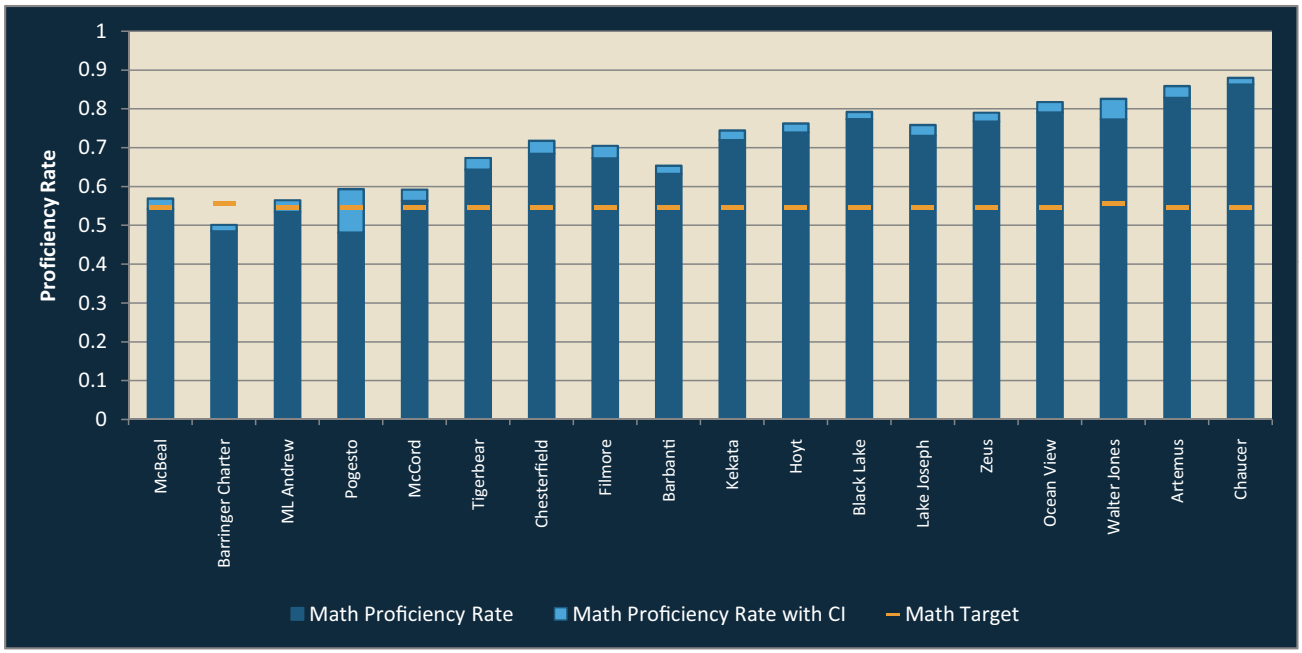

Figure 6. Impact of the confidence on middle school math proficiency rates for 2008s

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample middle schools (ML Andrew and Pogesto) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Nevada AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	37.9%	23.7%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Maryweather	46.1%	37.9%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	1
Few	51.2%	38.8%	N	N	N	N	N	N	N	N	Y	Y			Y	N			Y	N	14	4	29%	N	1
Nemo	51.2%	52.6%	Y	Y	N	N			N	N	N	N							Y	Y	10	4	40%	N	7
Island Grove	52.1%	56.4%	Y	Y	N	N	N	N	N	Y					N	N			Y	Y	12	5	42%	N	4
JFK	59.5%	46.6%	Y	N	N	N			Y	N	N	N							Y	Y	10	4	40%	N	3
Scholls	69.0%	56.1%	Y	Y	N	N	Y	N	Y	Y	N	N			Y	N			Y	Y	14	8	57%	N	7
Hissmore	68.1%	57.6%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	61.9%	58.9%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	5
Alice Mayberry	65.5%	57.4%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	9
Wayne Fine Arts	65.5%	67.8%	Y	Y					Y	Y	N	Y							Y	Y	8	7	88%	N	21
Winchester	67.5%	67.8%	Y	Y	N	N			Y	Y					Y	Y			Y	Y	10	8	80%	N	22
Coastal	74.4%	63.9%	Y	Y	N	N	N	N	Y	Y	Y	N			Y	Y			Y	Y	14	9	64%	N	3
Paramount	74.0%	66.8%	Y	Y	Y	Y	N	N	Y	N					Y	N			Y	Y	12	8	67%	N	7
Forest Lake	82.0%	76.1%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	86.3%	76.9%	Y	Y	Y	N	Y	N	Y	Y			Y	Y	Y	N			Y	Y	14	11	79%	N	10
Roosevelt	86.9%	83.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	28
King Richard	84.1%	82.7%	Y	Y	N	N	N	N	N	N					Y	Y			Y	Y	12	6	50%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

The effect of confidence intervals on reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, only one elementary school (John F. Kennedy) and one middle school (Chesterfield) met the overall targets with the confidence interval, although we know from Figures 3 and 4 that these two schools still failed to meet all of their subgroup targets. In short, **the application of the confidence interval had only modest effect on whether the sample elementary and middle schools met Nevada's overall reading and math targets.**[9]

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still pass AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to *show* how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 3. Middle school subgroup performance of sample schools under the 2008 Nevada AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	54.2%	51.1%	Y	N	N	N	N	N	N	N	N	N	Y	Y	N	N	Y	N	Y	Y	18	6	33%	N	0
Barringer Charter	48.5%	46.8%	N	N	N	N			N	N	N	N			Y	Y			Y	Y	12	4	33%	N	0
ML Andrew	53.5%	54.5%	Y	N	N	N	N	N	N	N	N	N			Y	N			Y	Y	14	4	29%	N	0
Pogesto	48.1%	57.4%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	15
McCord Charter	56.3%	59.4%	Y	Y	N	N	N		N	N	N	N			N	N			Y	Y	13	4	31%	N	0
Tigerbear	64.3%	53.7%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	0
Chesterfield	68.5%	56.3%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	1
Filmore	67.3%	64.0%	Y	Y	N	N	N	N	Y	Y					Y	N			Y	Y	12	7	58%	N	1
Barbanti	63.3%	59.9%	Y	Y	N	N	N	N	N	N					Y	N			Y	Y	12	5	42%	N	0
Kekata	71.9%	64.0%	Y	Y	N	N	N	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	0
Hoyt	73.9%	67.0%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	2
Black Lake	77.3%	67.5%	Y	Y	N	N	Y	N	Y	N	Y	N	Y	Y	Y	Y			Y	Y	16	11	69%	N	0
Lake Joseph	73.1%	70.6%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	76.8%	70.5%	Y	Y	N	N	N	N	Y	Y	Y	Y			N	N			Y	Y	14	8	57%	N	1
Ocean View	79.1%	79.7%	Y	Y	Y	N	N	N	Y	N			Y	Y	Y	N			Y	Y	14	9	64%	N	2
Walter Jones	77.3%	76.0%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	82.9%	77.1%	Y	Y	N	N			Y	N			Y	Y	N	N			Y	Y	12	7	58%	N	3
Chaucer	86.3%	85.3%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Nevada rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- The majority of schools met their targets for their overall student school populations, but failed to make AYP because of the performance of one or more subgroups.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Nevada AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	14	15
Students with limited English proficiency	10	8	10
Low-income students	18	7	9
African-American students	9	5	6
Asian/Pacific Islander students	1	0	0
Hispanic students	12	4	8
American Indian/Alaska Native students	0	0	0
White students	17	0	1

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Nevada AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	15	16
Students with limited English proficiency	11	10	10
Low-income students	18	5	11
African-American students	11	4	9
Asian/Pacific Islander students	5	0	0
Hispanic students	14	4	9
American Indian/Alaska Native students	1	0	1
White students	18	0	0

- Four elementary schools failed to meet the reading targets and three failed to meet the math targets for their overall school populations.
- Four middle schools failed to meet the reading targets and one (Barringer Charter) failed to meet the math target for its overall school population.
- Low-income students tended to perform better on their math targets than their reading targets especially at the middle school level .

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. The performance of SWDs is proving challenging for schools under Nevada's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum n of

Table 6. Comparisons between schools that did and didn't make AYP in Nevada, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	1	17	2	16
Average student body size	262	307	124	951
Average % low income	13	48	42	45
Average % nonwhite	19	42	27	46
Average performance†	8.85	0.78	0.40	-0.11
Average % growth‡	103	116	109	97
Average number of targets to meet	8	11	7	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

25. In fact, for SWDs, only one elementary school (Paramount) (and no middle schools) met its targets in reading, and only two elementary schools (Paramount and Marigold) and one middle school (Ocean View) met their targets in math. Students with limited English proficiency are also struggling to meet the state's targets; every school with a large enough LEP population to qualify as a separate subgroup failed to meet its reading targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Nevada's NCLB accountability system is, in some respects, behaving like those in other states. For example, Roosevelt and Walter Jones were among those schools that made AYP in the greatest number of states—28 and 20. And these schools made AYP in Nevada, too.

But Nevada is also home to a few anomalies. First, consider Winchester Elementary (see Figure 3). It made AYP in 22 of the 28 states in our sample, but not in Nevada. In examining Table 2, we can see that Winchester met the minimum numbers for the SWD subgroup. Many other states within the sample had no SWD subgroup because of their larger minimum subgroup size. With more accountable subgroups, Winchester didn't meet all its targets; hence it failed to make AYP. This is likely to be also true for Wayne Fine Arts Elementary, which failed for a single subgroup.

That fewer subgroups make it more likely to make AYP is consistent with the patterns shown in Table 6, which compares the schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations (especially at the middle school level), fewer subgroups (and thus fewer targets to meet), and much lower percentages of nonwhite students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under

Nevada's AYP rules (and AMOs) for 2008. We found that only 1 elementary school and 2 middle schools—3 in all, from a sample of 36—would have made AYP in Nevada.

Looking across the 28 state accountability systems examined in the study, this puts Nevada near the lower end of the sample distribution in terms of the number of schools making AYP (see Figure 1). We find that the number of elementary schools that made AYP in Nevada is exceeded in 26 other sample states. The high number of schools that didn't make AYP in Nevada is partly because Nevada's minimum subgroup size is relatively small compared to other states; this means more subgroups are held separately accountable for performance. In fact, a few sample schools that made AYP in most other states did not make it in Nevada, largely because of the state's *n* size. (This occurred despite Nevada's fairly low annual targets, which require barely half of students to be proficient in 2008).

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Nevada's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Most sample schools made AYP in Nevada for their student populations as a whole, without considering subgroups. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their students aren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over whether or not a school makes adequate yearly progress? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[10] doesn't the massive failure of these students to meet Nevada's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required

[10] See footnote 4.

AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

New Hampshire

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines New Hampshire's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough New Hampshire's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under New Hampshire's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against New Hampshire's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **14 of 18 elementary schools** and **17 of 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under New Hampshire's accountability system. (This high failure rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students.)
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools that made AYP in New Hampshire was exceeded in 12 other sample states. New Hampshire ties Maine and New Mexico with 4 elementary schools making AYP. In addition, New Hampshire is one of 6 states with just a single passing middle school in the sample (see Figure 1).**
- Many of the schools in our sample that failed to make AYP in New Hampshire met expected targets for their overall populations[2] but failed because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.

New Hampshire is squarely in the middle of the state distribution in terms of the number of schools making AYP. This is not surprising given New Hampshire's complex rule set. First, New Hampshire's 99 percent confidence interval provides schools with greater leniency than the more commonly used 95 percent confidence interval. Second, the state awards students "partial credit" for performing at lower levels of proficiency. On the other hand, New Hampshire's annual targets require that schools reach a relatively high bar (e.g., in 2008, 86 percent of students in all subgroups must reach proficiency on the state's reading exam in order to make AYP). So, while the state's definitions of proficiency generally ranked about average compared with the standards set by other states, getting 86 percent of all students over that bar is relatively difficult. Finally, New Hampshire's minimum subgroup size is 11, which is much smaller than the subgroup size in most other states we examined. This means that more subgroups are held separately accountable for performance than would be in other jurisdictions.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the New England Common Assessment Program.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

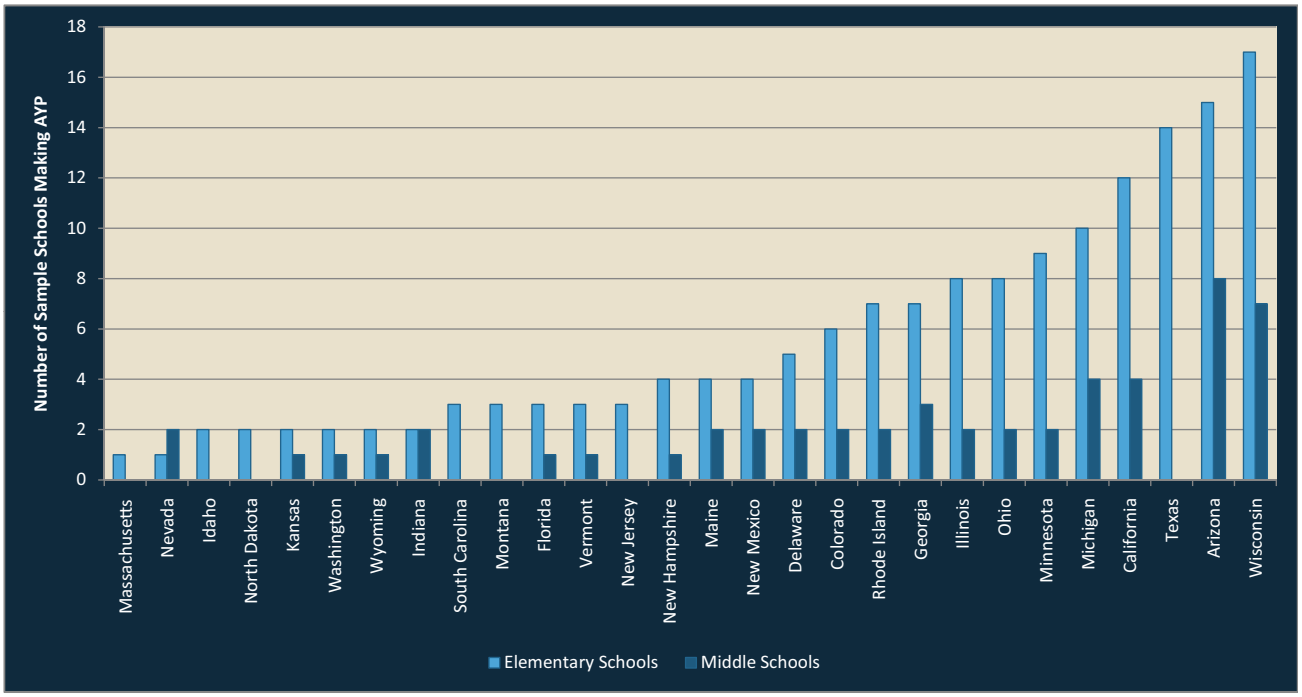

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- As in most states, middle schools in New Hampshire had greater difficulty reaching AYP than elementary schools, possibly because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower.

- A strong predictor of whether or not a school would make AYP under New Hampshire's system is whether it has enough limited English proficient (LEP)[3] students or SWDs to qualify as a separate subgroup. Most schools with a LEP or SWD subgroup failed to make AYP.[4]

- **Although New Hampshire awards "partial credit" to students performing at lower levels and uses a fairly lenient confidence interval (margin of statistical error), most schools still failed to make AYP, partly because of New Hampshire's small minimum *n* size (which makes more subgroups accountable) and partly because of New Hampshire's fairly high annual targets or AMOs.**

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on New Hampshire's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be slightly more negative than actual findings, mostly because of the differences in testing practices between the Measures of Academic Progress (MAP), the assessment we used in this study, and in the New England Common Assessment Program, the standardized state test. Specifically, the U.S. Department of Education has issued new NCLB 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.

profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than 10 pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the New England Common Assessment Program are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that New Hampshire's definitions of proficiency generally ranked about average compared with the standards set by

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

[6] We gave all schools in our sample pseudonyms in this report.

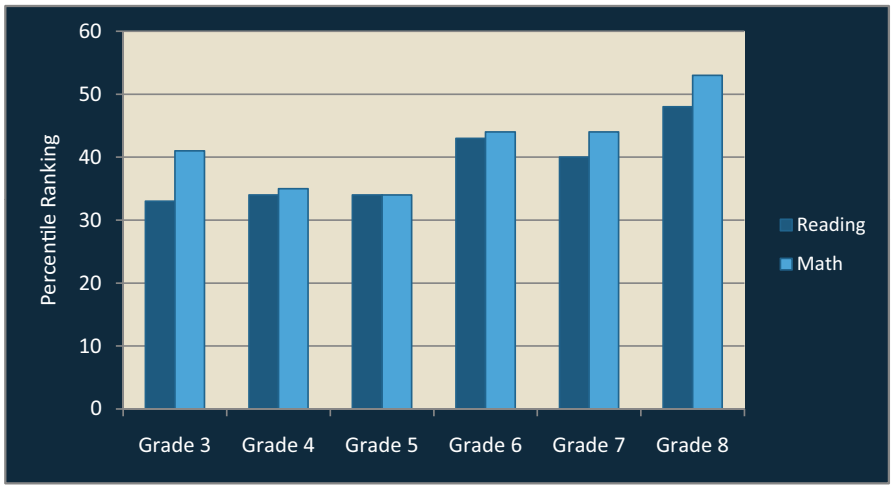

Figure 2. New Hampshire reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of New Hampshire's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of New Hampshire's cut scores are below the 55th percentile.

the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the New Hampshire test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under New Hampshire AYP rules for 2008. So to clarify, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against New Hampshire's 2008 AYP rules.

Table 1 shows the pertinent New Hampshire AYP rules that we applied to elementary and middle schools in the current study. New Hampshire's minimum subgroup size is 11, which is much smaller than the ones in most other states we examined.[7] **This means that schools in New Hampshire have more accountable subgroups than do similar schools in other states.**

Most states also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. **New Hampshire's 99% confidence interval, however, gives schools greater leniency than the more commonly used 95% confidence interval.** This means that if the annual target requires a school to achieve, for example, 86% reading proficiency among its grade 3-8 students (and 86% reading proficiency among its grade 3-8 students in each subgroup), applying the confidence interval means that the real target can be lower, particularly with smaller groups. Finally, rather than simply measuring the percentage of students achieving a "proficient" or higher performance level, **New Hampshire employs a proficiency "index," which gives partial credit to students performing at levels less than proficient.** In the short term, the index makes it easier for schools to achieve their targets, though as the targets approach the 100% requirement of NCLB in 2014, the assistance of the index diminishes.[8]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006),

[7] It's also likely that New Hampshire has small schools so a small *n* size may be appropriate.

[8] In six of the states studied (Massachusetts, Minnesota, Rhode Island, Vermont, and Wisconsin, as well as New Hampshire), an index is used that 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.

Table 1. New Hampshire AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 11	
	SWDs: 11	
	Low-income students: 11	
	LEP students: 11	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (index)**	**2008 targets (index)**
READING/LANGUAGE ARTS		
Grade 3	82	86
Grade 4	82	86
Grade 5	82	86
Grade 6	82	86
Grade 7	82	86
Grade 8	82	86
MATH		
Grade 3	76	82
Grade 4	76	82
Grade 5	76	82
Grade 6	76	82
Grade 7	76	82
Grade 8	76	82

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under New Hampshire's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under New Hampshire's 2008 AYP rules. **Only 4 elementary schools (Wayne Fine Arts, Win-**

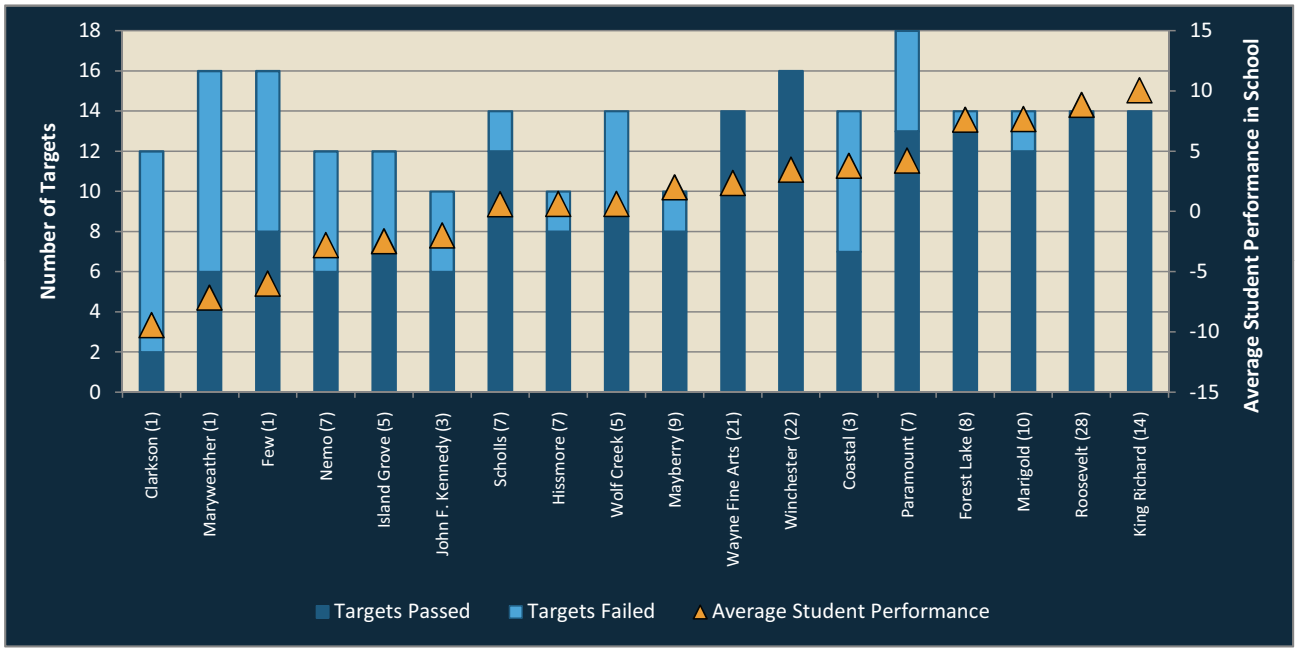

Figure 3. AYP performance of the elementary school sample under New Hampshire's 2008 AYP rules

Note: This figure shows how each of the elementary schools within the sample fared under New Hampshire's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Marigold Elementary, for example, met 12 of its 14 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

chester, Roosevelt, and King Richard) made AYP and 14 failed. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. All schools that made AYP are in the right half of the figure, meaning that relatively high performing students were found at these schools.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 New Hampshire AYP rules. **Of 18 middle schools in our sample, only 1 made AYP**—a high-performance school (Walter Jones) that has relatively few qualifying subgroups compared to other schools.

Figures 5 and 6 indicate the degree to which math proficiency rates are aided by New Hampshire's confidence interval for elementary and middle schools, respectively. On these figures, the darker portion of the bars show the actual proficiency rates at each school, and the lighter portion of the bars show the degree to which these proficiency rates are increased by the application of the confidence interval. The orange lines show the AMO needed to meet AYP. These figures show that four elementary schools (Few, Island Grove, Nemo, and Wolf Creek) and two middle schools (Hoyt and Lake Joseph) were assisted by the confidence intervals to meet their overall targets in math (note how the orange line falls within the light blue band); all of these schools, however, still failed to make AYP because of low subgroup performance (see Figures 3 and 4).

The effect of the confidence intervals on reading proficiency rates at the elementary and middle school levels is much the same (not shown). In reading, six elementary schools (Nemo, Island Grove, JFK, Scholls, Wolf Creek, and Coastal) and two middle schools (Pogesto and Lake Joseph) met their overall targets with the help of the confidence interval. However, we know from Fig-

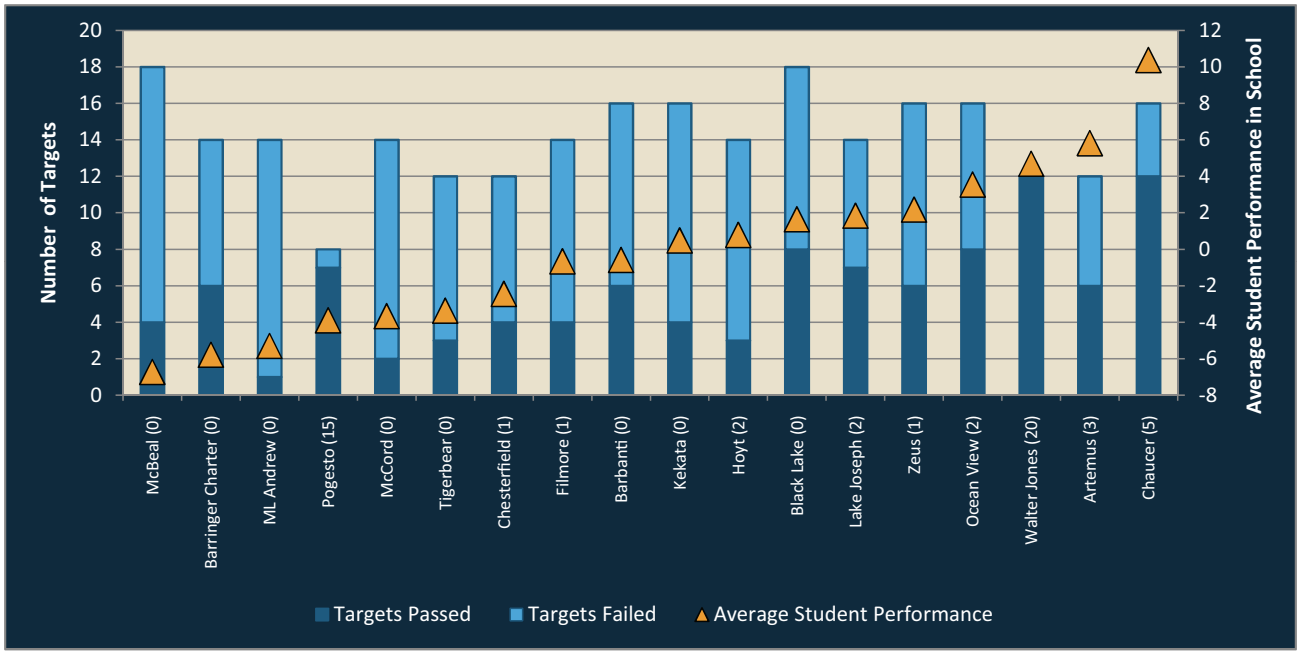

Figure 4. AYP performance of the middle school sample under New Hampshire's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under New Hampshire's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Pogesto, for example, met 7 of its 8 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

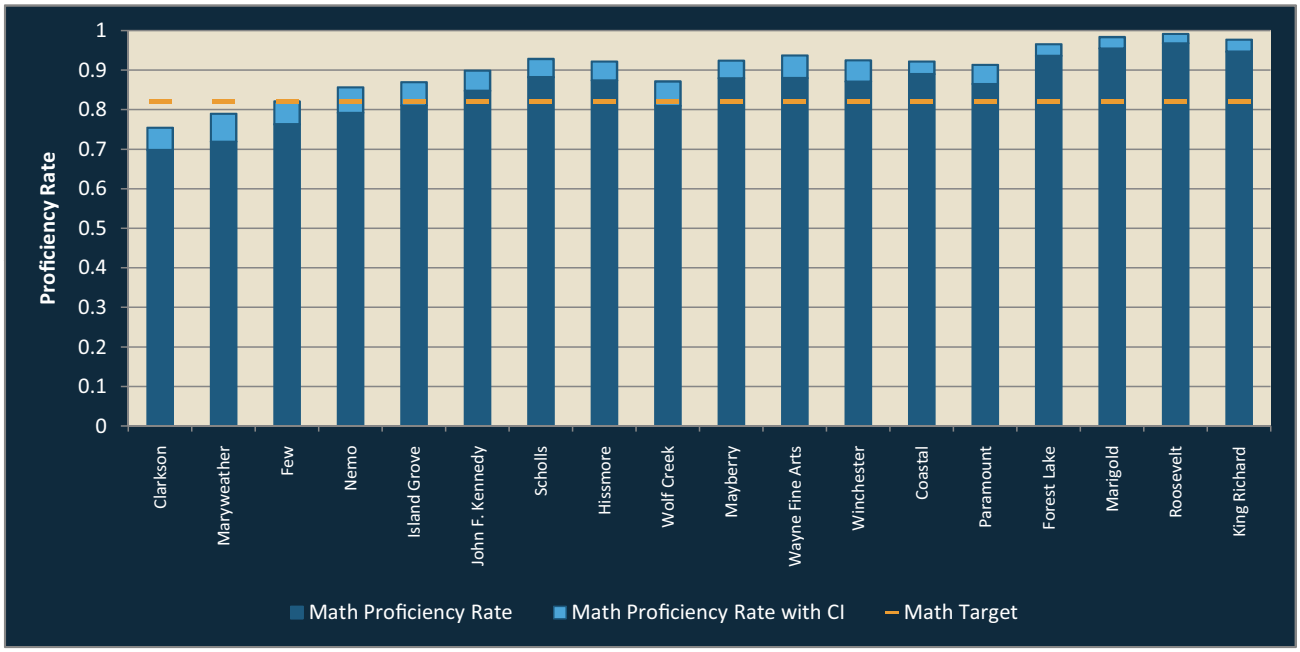

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under New Hampshire's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that four of the elementary schools (Few, Island Grove, Nemo, and Wolf Creek) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

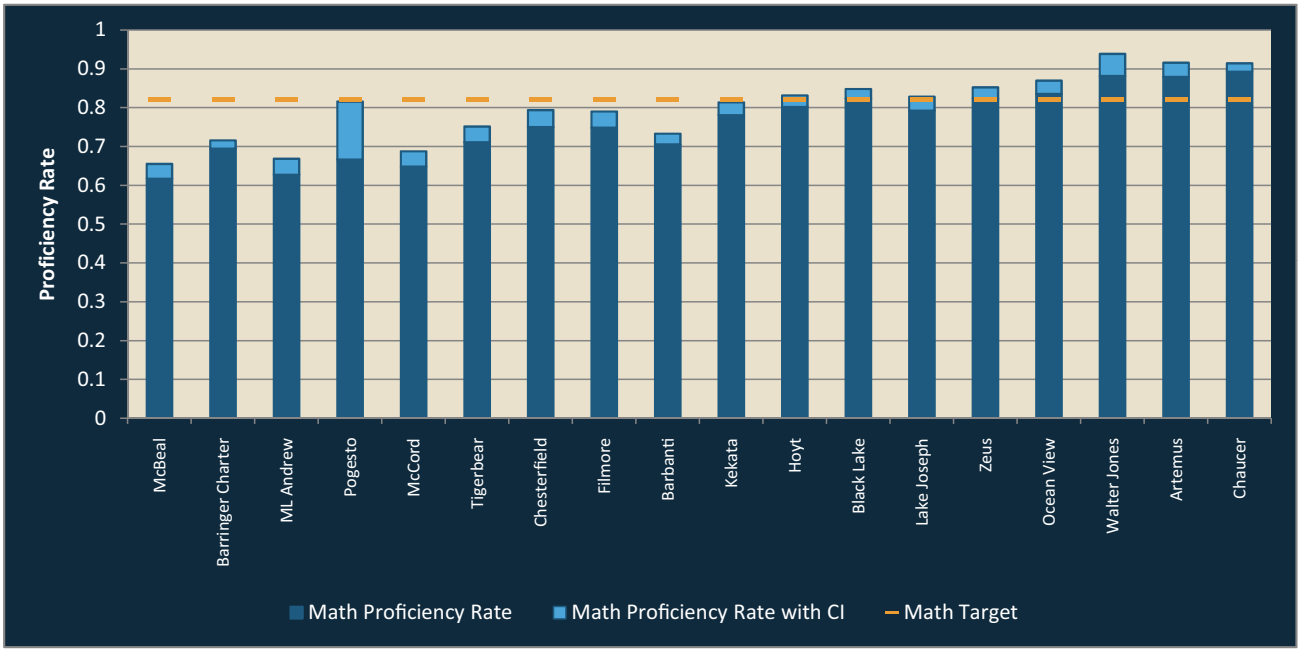

Figure 6. Impact of the confidence interval on middle school math proficiency rates under New Hampshire's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that two of the sample middle schools (Hoyt and Lake Joseph) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

ures 3 and 4 that all these schools failed to meet their targets for some subgroups. **Overall, the application of the confidence interval, despite the fact that it is lenient, seems to have little or no effect on AYP outcomes for the sample elementary and middle schools in New Hampshire.**[9]

Where Do Schools Fail?

Figures 3 and 4 illustrate the number of subgroup targets at the sample elementary and middle schools and the number of targets met in New Hampshire. However, these figures do not indicate which subgroups passed or failed in each school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 New Hampshire rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Only two elementary schools (Clarkson and Maryweather) failed to meet both the reading and the math targets for their overall school population.
- About half of the middle schools failed in both reading and math for their overall student populations.

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary school subgroup performance of sample schools under the 2008 New Hampshire AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	69.8%	66.5%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	1
Maryweather	72.0%	69.6%	N	N	N	N	N	N	N	N	Y	Y			N	N	Y	Y	Y	Y	16	6	38%	N	1
Few	76.4%	72.9%	Y	N	N	N	N	N	N	N	Y	Y			Y	N	Y	Y	Y	Y	16	8	50%	N	1
Nemo	79.3%	83.7%	Y	Y	N	N			N	N	N	N			Y	Y			Y	Y	12	6	50%	N	7
Island Grove	81.1%	82.2%	Y	Y	N	N	N	N	Y	Y					Y	N			Y	Y	12	7	58%	N	4
JFK	84.8%	81.1%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	3
Scholls	88.3%	84.2%	Y	Y	N	N	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	12	86%	N	7
Hissmore	87.5%	86.3%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	81.0%	83.6%	Y	Y	N	N	N	N	Y	Y			Y	Y	N	Y			Y	Y	14	9	64%	N	5
Alice Mayberry	88.0%	88.3%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	9
Wayne Fine Arts	88.0%	93.9%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	21
Winchester	87.2%	90.1%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	16	100%	Y	22
Coastal	89.1%	85.1%	Y	Y	N	N	N	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	3
Paramount	86.5%	86.5%	Y	Y	N	Y	N	N	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	18	13	72%	N	7
Forest Lake	93.7%	93.3%	Y	Y	Y	N			Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	14	13	93%	N	8
Marigold	95.5%	92.5%	Y	Y	Y	Y	Y	N	Y	Y			Y	Y	Y	N			Y	Y	14	12	86%	N	10
Roosevelt	96.8%	96.9%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	28
King Richard	94.7%	94.5%	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y			Y	Y	14	14	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

- Four elementary schools (Scholls, Hissmore, Alice Mayberry, and Forest Lake) met every target except for their SWDs.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. We see that the performance of SWDs is proving very challenging for schools under New Hampshire's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 11. The same is true for students with limited English proficiency. In fact, all but one middle school (Walter Jones) in the study with qualifying SWD and two middle schools (Barringer Charter and McCord Charter) with qualifying LEP subgroups failed to meet their targets for these subgroups in reading or math. Low-income students are also struggling to meet the state's targets. Most middle schools with a large enough low-income population to qualify as a separate subgroup failed to meet their reading and math targets for these students (recall that proficiency cut scores in math and reading are generally lower at the elementary than the middle school level).

Other state reports contain a section comparing some of

Table 3. Middle school subgroup performance of sample schools under the 2008 New Hampshire AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	61.8%	68.7%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	69.4%	76.9%	N	N	N	N	Y	Y	N	N	N	N			Y	Y			Y	Y	14	6	43%	N	0
ML Andrew	62.8%	75.6%	N	N	N	N	N	N	N	N	N	N			N	N			N	Y	14	1	7%	N	0
Pogesto	66.7%	78.9%	N	Y					Y	Y					Y	Y			Y	Y	8	7	88%	N	15
McCord Charter	64.8%	77.8%	N	N	N	N	N	Y	N	N	N	N			N	N			N	Y	14	2	14%	N	0
Tigerbear	71.1%	72.6%	N	N	N	N			N	N	N	N			Y	N			Y	Y	12	3	25%	N	0
Chesterfield	75.0%	76.8%	N	N	N	N			N	N	N	N			Y	Y			Y	Y	12	4	33%	N	1
Filmore	74.9%	82.0%	N	N	N	N	N	N	N	N			Y	Y	N	N			Y	Y	14	4	29%	N	1
Barbanti	70.5%	77.3%	N	N	N	N	N	N	N	N	Y	Y	Y	Y	N	N			Y	Y	16	6	38%	N	0
Kekata	78.1%	79.7%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N			Y	Y	16	4	25%	N	0
Hoyt	80.2%	82.1%	Y	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	3	21%	N	2
Black Lake	82.4%	81.8%	Y	N	N	N	N	N	N	N	N	N	Y	Y	Y	N	Y	Y	Y	Y	18	8	44%	N	0
Lake Joseph	79.3%	84.8%	Y	Y	N	N	N	N	N	Y	Y	Y			N	N			Y	Y	14	7	50%	N	2
Zeus	82.4%	83.1%	Y	N	N	N	N	N	N	N	Y	N	Y	Y	N	N			Y	Y	16	6	38%	N	1
Ocean View	83.5%	89.3%	Y	Y	N	N	N	N	N	N	Y	Y	Y	Y	N	N			Y	Y	16	8	50%	N	2
Walter Jones	88.1%	89.9%	Y	Y	Y	Y			Y	Y	Y	Y			Y	Y			Y	Y	12	12	100%	Y	20
Artemus	87.9%	87.7%	Y	Y	N	N			N	N			Y	Y	N	N			Y	Y	12	6	50%	N	3
Chaucer	89.3%	92.5%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

the characteristics of the sample schools that made AYP versus those that did not. In New Hampshire, there were no striking differences between schools that made AYP and those that didn't, either at the elementary or middle school level. The one exception (rather expected) was that schools that made AYP had students with higher average performance than did schools that didn't make it, as measured by NWEA reading and math tests.[10]

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under New Hampshire's AYP rules (and AMOs) for 2008. We found that only 4 elementary schools and 1 middle school—just 5 out of a sample of 36—would have made AYP in New Hampshire. Looking across the 28 state ac-

[10] There were also no "anomalies" in New Hampshire. All the sample schools that made AYP in New Hampshire made it in the other states examined; similarly, sample schools that failed to make AYP in New Hampshire tended to fail in most other states as well.

Table 4. Summary of subgroup performance of sample elementary schools under 2008 New Hampshire AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	18	12	12
Students with limited English proficiency	13	7	8
Low-income students	18	4	7
African-American students	13	1	3
Asian/Pacific Islander students	6	0	0
Hispanic students	15	3	7
American Indian/Alaska Native students	3	0	0
White students	18	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 New Hampshire AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	17	16	16
Students with limited English proficiency	13	12	11
Low-income students	18	15	14
African-American students	15	9	10
Asian/Pacific Islander students	9	0	0
Hispanic students	18	11	13
American Indian/Alaska Native students	2	1	1
White students	18	2	0

countability systems examined in the study, this puts New Hampshire roughly in the middle of the sample distribution in terms of the number of schools making AYP (see Figure 1). So, although New Hampshire awards "partial credit" to students performing at lower levels and uses a fairly lenient confidence interval (margin of error), most schools still failed to make AYP, partly because New Hampshire's small minimum *n* size (which makes more subgroups accountable) and partly because of New Hampshire's fairly high annual targets or AMOs.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, New Hampshire's NCLB account-

ability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or minority youngsters. Some of the sample schools met the New Hampshire reading and math targets for their student populations as a whole, that is, without considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their students aren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Is it "fair" that, in New Hampshire and in a handful of other states, students are awarded "partial" credit even though they do not achieve proficiency? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[11] doesn't the massive failure of these students to meet New Hampshire's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other pupils, but when almost no school is able to meet the goal perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[11] See footnote 4.

New Jersey

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 towards these goals.

This report examines New Jersey's NCLB accountability system—particularly how its various rules, criteria and practices result in schools either making AYP—or not making AYP. It also gauges how tough New Jersey's system is compared with other states. We selected 36 schools from around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether or not each would make AYP under New Jersey's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against New Jersey's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **15 of 18 elementary schools** in our sample **failed to make AYP** in 2008 under New Jersey's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students. **It's also likely due to New Jersey's low minimum *n* size of 20 (for most subgroups) and its fairly high annual targets, especially in reading.**
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in New Jersey was exceeded by 15 other sample states (New Jersey ties with 4 other states that each have 3 elementary schools making AYP).** This puts New Jersey in the lower part of the sample distribution (see Figure 1). (Note that middle schools were not examined in New Jersey, unlike other states, since eighth grade cut scores were not available.)
- Most of the schools in our sample that fail to make AYP in New Jersey are meeting expected targets for their overall populations but failing because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[2]
- As is the case in other states, schools with fewer subgroups attain AYP more easily in New Jersey than schools with more subgroups, even when their average student performance is lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

New Jersey falls near the middle of the state distribution in terms of the number of schools that make AYP. One particularly interesting thing about New Jersey is that a large group of Hispanic/Latino, African American, and low-income students met their targets in math. This is unusual because New Jersey's minimum subgroup size for these groups (20) is smaller than most other states', meaning that schools in New Jersey are held accountable for more subgroups than would similar schools in other states. However, New Jersey's definitions of proficiency generally ranked below average compared with the standards set by the other states, especially in grades 3-5 math. This likely accounts for the higher pass rate for traditionally disadvantaged groups.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the New Jersey Assessment of Skills and Knowledge (NJ ASK).

[2] SWDs are defined as those students following individualized education plans.

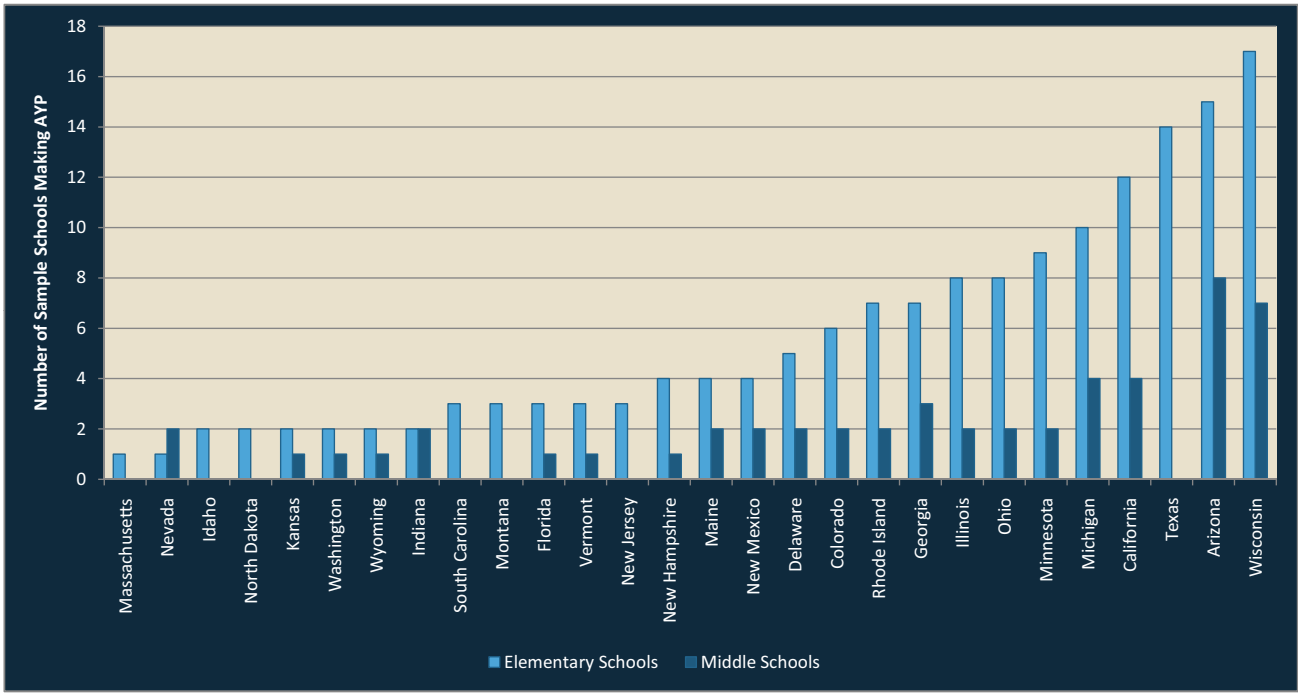

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- A strong predictor of a school making AYP under New Jersey's system is whether it has enough English language learners and SWDs to qualify as separate subgroups. Every single elementary school with limited English proficient and SWD subgroups failed to make AYP, in part because these students did not meet the state's targets in reading.[3]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on New Jersey's tests and 25 other state tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* to examine other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimate how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-perform-

[3] It should be noted that our subgroup findings for Limited English Proficient (LEP) and students with disabilities may be slightly more negative than would be seen under real world conditions. This is mostly due to the differences in testing practices between how LEP students and students with disabilities are treated in the NWEA's Measures of Academic Progress (MAP), the assessment used in this study, and in the New Jersey Assessment of Skills and Knowledge (NJ ASK), the state standardized assessment. Specifically, the U.S. Department of Education has issued NCLB guidelines permitting schools to exclude small percentages of LEP or disabled students from taking state tests, or providing them alternate assessments. In the current study, however, no valid MAP scores were omitted from consideration.

ing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by school year 2013–14. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (low income[4] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could both jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) As a result of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as ten youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state. This means that a school making AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005-06 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group such as English language learners,[5] among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the New Jersey Assessment of Skills and Knowledge (NJ ASK) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that New Jersey's definitions of proficiency generally ranked below average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the New Jersey test, given their performance on MAP. Student test data and subgroup designations are then used to determine how these 18 elementary schools would have fared under New Jersey AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–06, but we are applying them against New Jersey's 2008 AYP rules. Note that in New Jersey, unlike most of the other state reports, the 18 sample middle schools were not examined since New Jersey's eighth grade cut scores were not available.

Table 1 shows the pertinent New Jersey AYP rules that were applied to elementary schools in the current study.

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

[5] Note that we use "students with limited English proficiency (LEP)" or "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup.

[6] We gave all schools in our sample pseudonyms in this report.

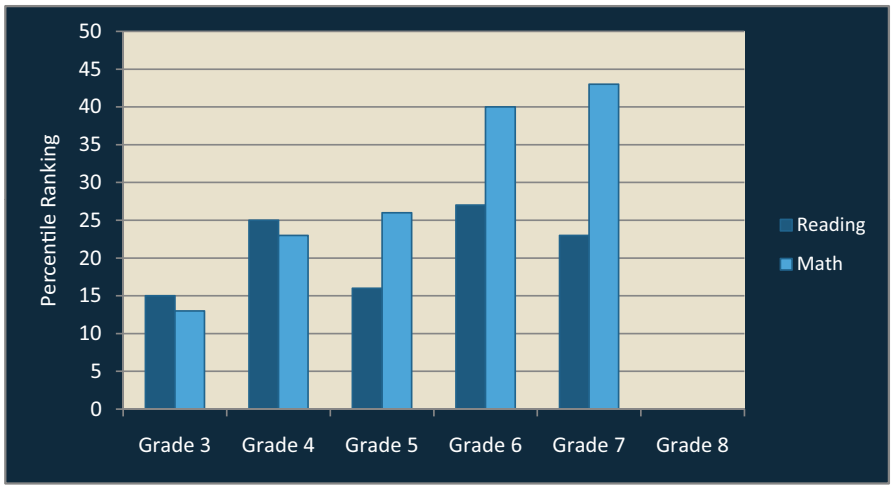

Figure 2. New Jersey reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of New Jersey's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through seven. Cut scores were not available for grade eight. Higher percentile ranks are more difficult to achieve. All of New Jersey's cut scores are below the 45th percentile. Cut score estimates for 8th grade were not available.

New Jersey's minimum subgroup size is 20 for all groups except for SWDs which is 35. While 35 is fairly consistent with the sizes used by most other states, 20 is smaller than most.[7] This means that schools in New Jersey will be accountable for more subgroups than would similar schools in other states.

Most states also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. The 95% confidence interval applied to proficiency rate calculations in New Jersey is comparable to the majority of states examined in the study. So, for instance, though schools are supposed to get 82% of their grade 3 students (as well as 82% of their students in each subgroup) to the proficient level on the state reading test, applying the confidence interval means that the real target can actually be lower, particularly with smaller groups.

Note that we were not able to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005-2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each school's subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

[7] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

Table 1. New Jersey AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 20	
	SWDs: 35	
	Low-income students: 20	
	LEP students: 20	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	82
Grade 4	68	82
Grade 5	n/a	82
Grade 6	n/a	76
Grade 7	n/a	76
Grade 8	58	76
MATH		
Grade 3	n/a	73
Grade 4	53	73
Grade 5	n/a	73
Grade 6	n/a	62
Grade 7	n/a	62
Grade 8	39	62

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not available

How Did the Sample Schools Fare Under New Jersey's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under New Jersey's 2008 AYP rules. **Only three elementary schools made AYP (Wayne Fine Arts, Winchester, and Roosevelt) while fifteen did not**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. All schools that made AYP are in the right half of the figure, meaning that the higher performing students were found at these schools.

Yet among these high performing schools, the only schools actually to make AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has separate targets). For example, Winchester passed, but has only nine targets. Among the eighteen elementary schools, this school has the fewest subgroups in New Jersey (along with Clarkson).

Figures 4 and 5 indicate the degree to which elementary schools' reading and math proficiency rates are aided by New Jersey's confidence interval. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates were increased by applying the

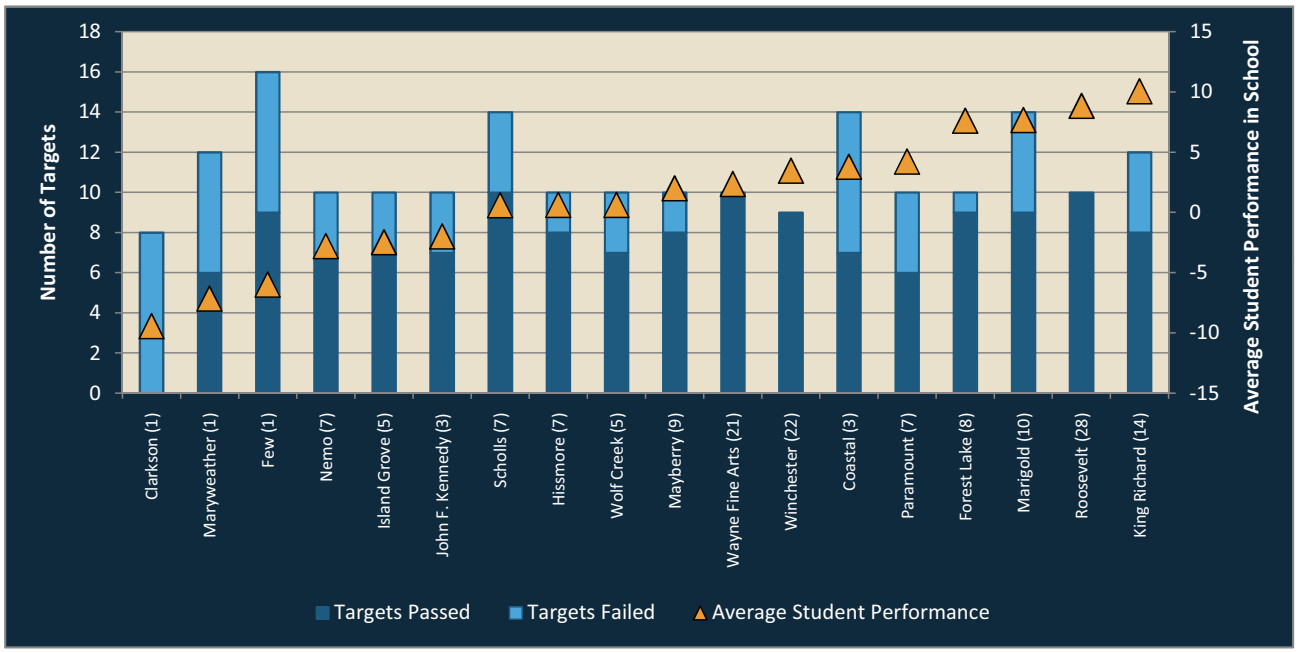

Figure 3. Performance of the elementary school sample under New Jersey's 2008 AYP Rules

Note: This figure indicates how each of the elementary schools within the sample fared under the New Jersey AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Forest Lake, for example, meets nine of its ten targets but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

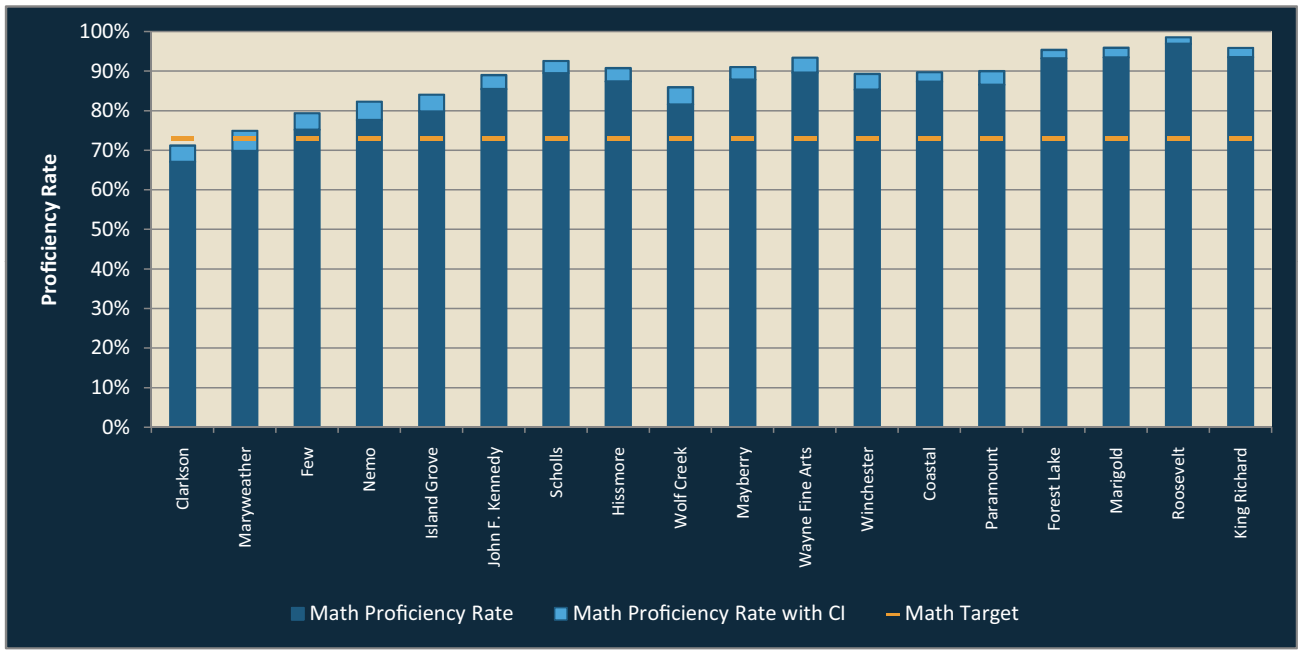

Figure 4. Impact of the confidence interval on elementary school math proficiency rates for 2008

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (Maryweather) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

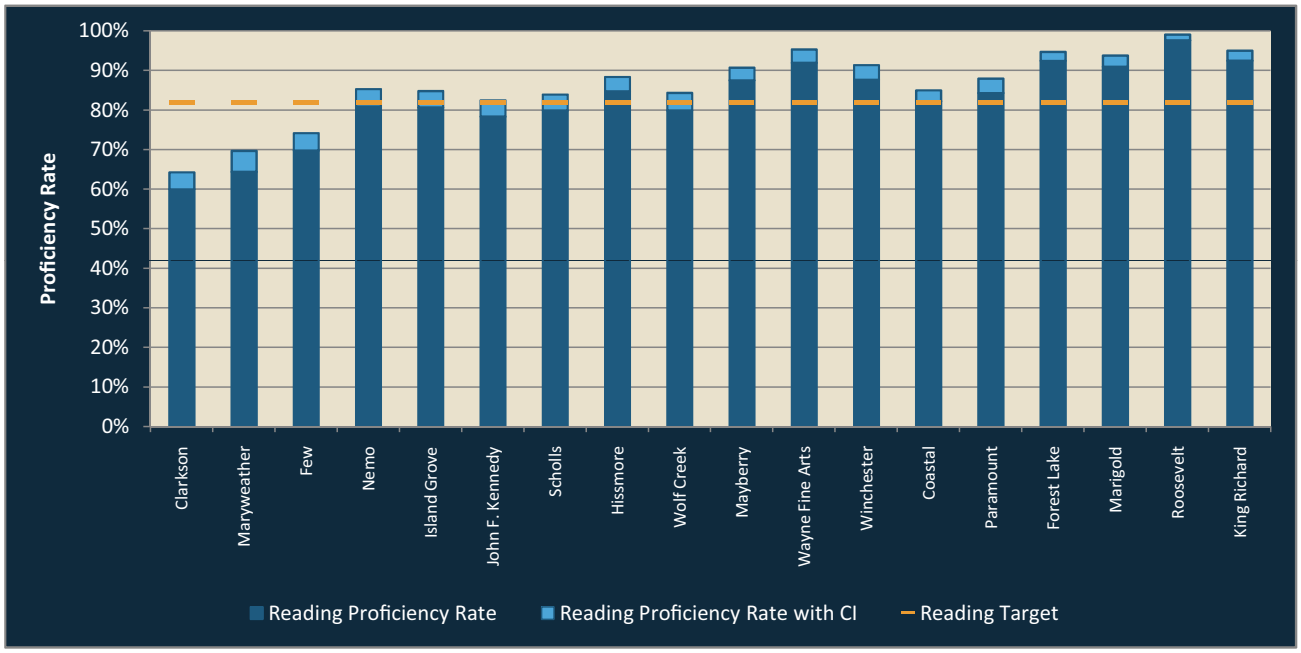

Figure 5. Impact of the confidence interval on elementary school reading proficiency rates for 2008

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that four of the sample elementary schools (Nemo, Island Grove, Scholls, and Wolf Creek) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

confidence interval. The orange lines show the annual measurable objective needed to meet AYP. Figure 4 shows that one of the sample elementary schools (Maryweather) met its overall math target with the assistance of the confidence interval (note how the orange bar falls in the light blue band). In reading (Figure 5), four schools (Nemo, Island Grove, Scholls, and Wolf Creek) were able to achieve their overall targets when assisted by the confidence interval. All of these schools, however, still fail to make AYP because of low subgroup performance (shown in Figure 3). **Overall, the application of the confidence interval had no effect on whether the sample schools met their overall reading or math targets in New Jersey.**[8]

Where do schools fail?

Figure 3 illustrates how the number of subgroups can impact the AYP decisions for our sample schools, but it conveys no information about which subgroups failed or passed in which school. Table 2 lists information on individual subgroup performance.

Table 2 shows which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP purposes include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Table 2 also shows whether a school made AYP under the New Jersey rules, and the total number of states within the study in which that school met AYP.

[8] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 4 and 5. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary subgroup performance of sample schools under the 2008 New Jersey AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	67.1%	60.0%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	69.9%	64.4%	Y	N			N	N	Y	N	Y	Y			N	N			Y	Y	12	6	50%	N	1
Few	75.3%	69.8%	Y	N	N	N	N	N	Y	N	Y	Y			Y	N	Y	Y	Y	Y	16	9	56%	N	1
Nemo	77.7%	80.9%	Y	Y					Y	N	N	N			Y	Y			Y	Y	10	7	70%	N	7
Island Grove	79.8%	80.7%	Y	Y			N	N	Y	Y					Y	N			Y	Y	10	7	70%	N	4
JFK	85.5%	78.4%	Y	Y	Y	N			Y	N	Y	N							Y	Y	10	7	70%	N	3
Scholls	89.6%	79.9%	Y	Y	Y	N	Y	N	Y	N	Y	N			Y	Y			Y	Y	14	10	71%	N	7
Hissmore	87.5%	84.7%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	81.7%	79.9%	Y	Y			Y	N	Y	N					Y	N			Y	Y	10	7	70%	N	5
Alice Mayberry	87.9%	87.5%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	9
Wayne Fine Arts	89.7%	92.0%	Y	Y					Y	Y	Y	Y			Y	Y			Y	Y	10	10	100%	Y	21
Winchester	85.4%	87.7%	Y	Y					Y	Y				Y	Y	Y			Y	Y	9	9	100%	Y	22
Coastal	87.4%	82.3%	Y	Y	N	N	N	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	3
Paramount	86.6%	84.3%	Y	Y			N	N	Y	N					Y	N			Y	Y	10	6	60%	N	7
Forest Lake	93.3%	92.5%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	8
Marigold	93.5%	91.0%	Y	Y	Y	N	Y	N	Y	N			Y	Y	N	N			Y	Y	14	9	64%	N	10
Roosevelt	97.0%	97.6%	Y	Y					Y	Y	Y	Y			Y	Y			Y	Y	10	10	100%	Y	28
King Richard	93.6%	92.5%	Y	Y	Y	N	Y	N	Y	N					Y	N			Y	Y	12	8	67%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

The school-by-school findings in Tables 2 show that:

- Three elementary schools (Clarkson, Maryweather, and Few) failed to meet the reading targets for their overall school population. Only one elementary school (Clarkson) failed to meet its overall target in math.
- Three elementary schools (Hissmore, Alice Mayberry, and Forest Lake) met all their reading and math targets for all subgroups except for their SWDs.
- Most low-income students met their math but not their reading targets (perhaps because reading cut scores are generally higher than math in the lower grades, as are annual targets in reading).

Table 3 summarizes the performance of the various subgroups. As shown, the performance of SWDs is particularly challenging within our sample schools. Every school within the sample with sufficient numbers of students with disabilities to qualify as a subgroup failed to meet its reading targets (this was also true for students with limited English proficiency.)

Table 3. Summary of subgroup performance of sample elementary schools under the 2008 New Jersey AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	9	4	9
Students with limited English proficiency	10	6	10
Low-income students	18	1	11
African-American students	11	1	4
Asian/Pacific Islander students	1	0	0
Hispanic students	14	3	9
American Indian/Alaska Native students	1	0	0
White students	17	0	0

Other state reports contain a section comparing some of the characteristics of the sample schools that made AYP versus those that did not. In New Jersey, there were no striking differences between schools that did and didn't make AYP at the elementary level, other than the (expected) finding that the former had students with higher average student performance than the latter, as measured by NWEA reading and math tests.

Concluding Observations

This study examined the test performance data of students from 18 elementary schools across the country to see how they would fare under New Jersey's AYP rules (and AMOs) for 2008. We found that only three elementary schools would have made AYP in New Jersey. Looking across the 28 state accountability systems examined in the study, this puts New Jersey in the lower middle of the sample distribution in terms of schools making AYP (see Figure 1). Part of this may be due to New Jersey's low minimum *n* of 20 (for non-SWD subgroups) and its fairly high annual performance targets, especially in reading.

The overriding goal of the No Child Left Behind act (NCLB) is to eliminate educational disparities within and across states; it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, New Jersey's No Child Left Behind accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Many of the sample schools make AYP in New Jersey for their student populations as a whole, i.e., without considering subgroup results. In the pre-NCLB era, such schools might have been considered effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and students with disabilities are more generous under the current state assessment system,[9] doesn't the mas-

[9] See footnote 3.

sive failure of these students to meet New Jersey's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

New Mexico

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines New Mexico's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough New Mexico's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under New Mexico's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against New Mexico's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- **We estimate that 14 of 18 elementary schools and 16 of 18 middle schools in our sample failed to make AYP in 2008 under New Mexico's accountability system.** This high failure rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students. But it's also partly explained by New Mexico's minimum *n* size for subgroups, which tends to be smaller than those used in most other states, meaning it holds more subgroups accountable for performance.[2]
- The smaller *n* size appears to be a factor in the number of schools making AYP in New Mexico, despite the state's low overall cut scores in reading and low annual proficiency targets in math and reading (e.g., the state demands that only 35% of students in grades six through eight reach math proficiency in 2008).
- Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in New Mexico is exceeded in 12 other sample states (New Mexico ties with New Hampshire and Maine, each with 4 elementary schools making AYP). New Mexico is one of 10 states with 2 middle schools each that made AYP in the sample (see Figure 1).

There are some interesting dynamics that place **New Mexico** near the middle of the state distribution in terms of the number of schools making AYP. This is a state which has several rigorous requirements combined with more lenient ones. For example, New Mexico's cut scores in math are close to or above the 50th percentile, while reading cut scores mostly hover around the 30th percentile. So more rigor in math is coupled with less rigor in reading. New Mexico's 99 percent confidence interval provides schools with greater leniency than the more commonly used 95 percent confidence interval found in other states. However, New Mexico's minimum subgroup size is 25, which is smaller than most other states we examined. This means that schools in New Mexico will have more accountable subgroups than would similar schools in other states, making it difficult for large schools with many accountable subgroups to make AYP there.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the New Mexico Standards Based Assessments.

[2] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

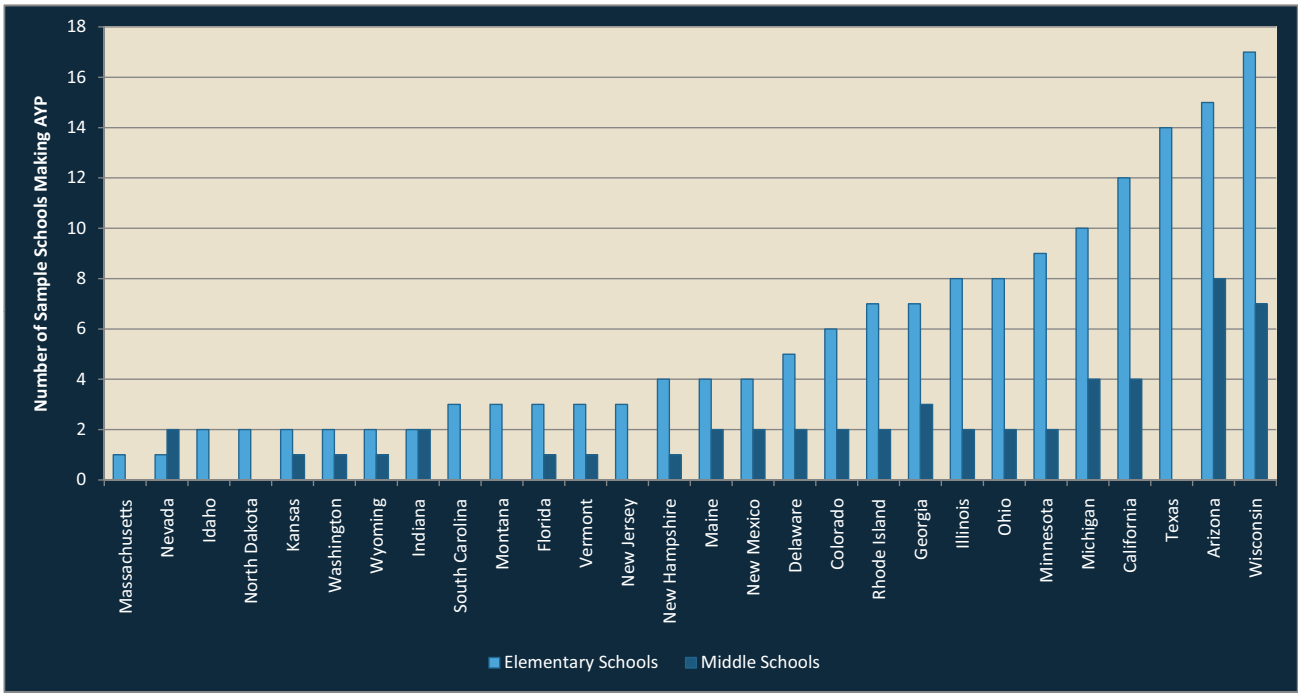

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- Nearly all of the schools in our sample that failed to make AYP in New Mexico are meeting expected targets for their overall populations[3] but failed because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.

- As in other states, middle schools in New Mexico had greater difficulty reaching AYP than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- Middle schools with fewer subgroups attained AYP more easily in New Mexico than middle schools with more subgroups, even when their average student performance is lower. In other words, schools with greater diversity and size face greater challenges in making AYP. This is the case in other states as well.

- A strong predictor of whether or not a school makes AYP under New Mexico's system is whether it has enough English language learners to qualify as a separate subgroup. Every single school with a limited English proficient (LEP)[4] subgroup failed to make AYP. Likewise, most of the schools (especially at the middle school level) with enough qualifying SWDs failed to meet their AYP targets.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on New Mexico's tests and those of 25

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup.

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

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 New Mexico Standards Based Assessments, the standardized state test. 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.

other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than 10 pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

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

[7] We gave all schools in our sample pseudonyms in this report.

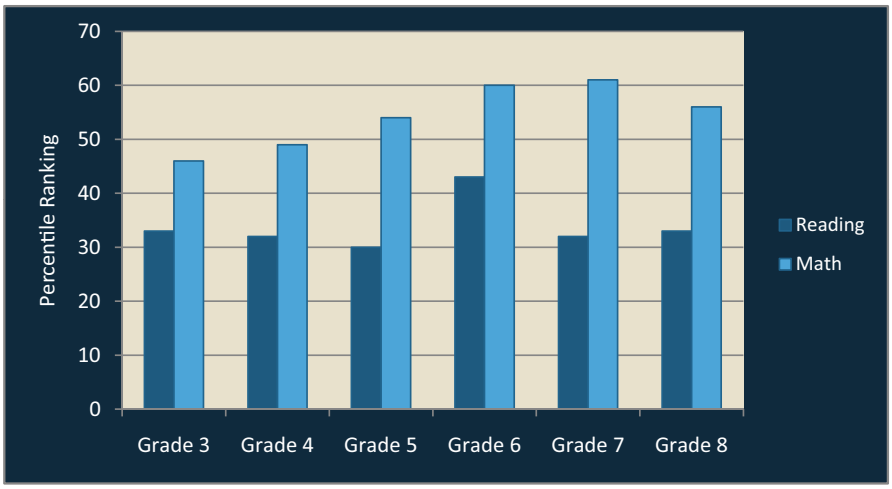

Figure 2. New Mexico reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of New Mexico's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of New Mexico's cut scores in reading are below the 50th percentile, but the cut scores in math are close to or above the 50th percentile.

Table 1. New Mexico AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 25	
	SWDs: 25	
	Low-income students: 25	
	LEP students: 25	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	59
Grade 4	30	59
Grade 5	n/a	59
Grade 6	n/a	53
Grade 7	n/a	53
Grade 8	39	53
MATH		
Grade 3	n/a	44
Grade 4	35	44
Grade 5	n/a	44
Grade 6	n/a	35
Grade 7	n/a	35
Grade 8	33	35

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).
Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not applicable

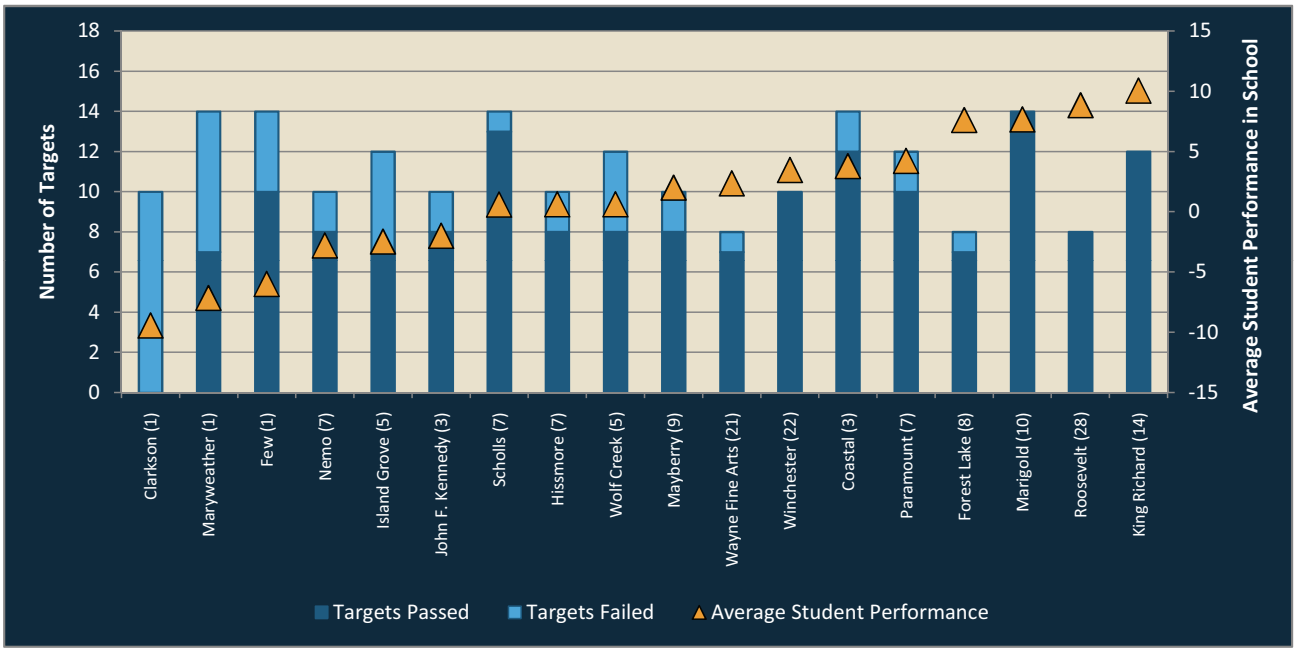

Figure 3. AYP performance of the elementary school sample under New Mexico's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under New Mexico's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Forest Lake, for example, met 7 of its 8 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

Proficiency cut score estimates for the New Mexico Standards Based Assessments (NMSBA) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that New Mexico's definitions of proficiency generally ranked below average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the New Mexico test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under New Mexico AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against New Mexico's 2008 AYP rules.

Table 1 shows the pertinent New Mexico AYP rules that we applied to elementary and middle schools in the current study. New Mexico's minimum subgroup size is 25, which is smaller than most other states we examined. This means that schools in New Mexico will have more accountable subgroups than would similar schools in other states.

Further, although most states also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates, New Mexico's 99% confidence interval gives schools greater leniency than the more commonly used 95% confidence interval. So, for instance, although schools are supposed to get 59% of their grade 3 students (and 59% of grade 3 students in each subgroup) to the proficient level on the state reading test, applying the confidence interval means that the real target can be lower, particularly with smaller groups.[8]

[8] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

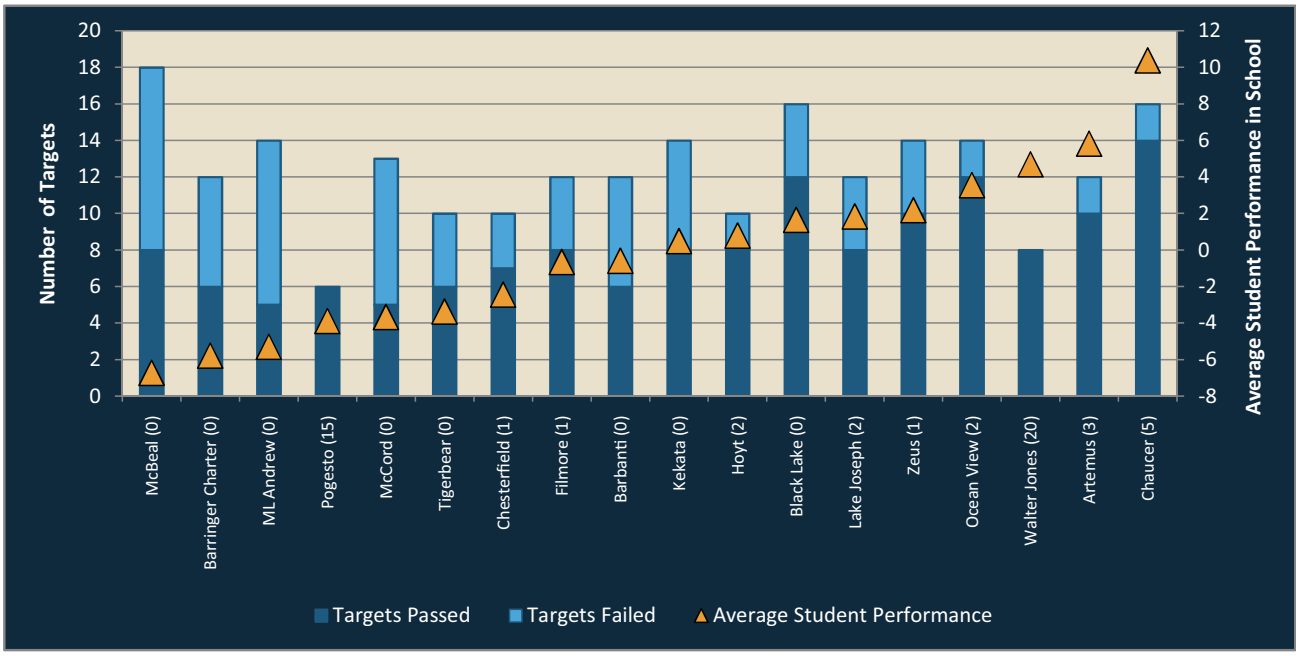

Figure 4. AYP performance of the middle school sample under New Mexico's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under New Mexico's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Artemus, for example, met 10 of its 12 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

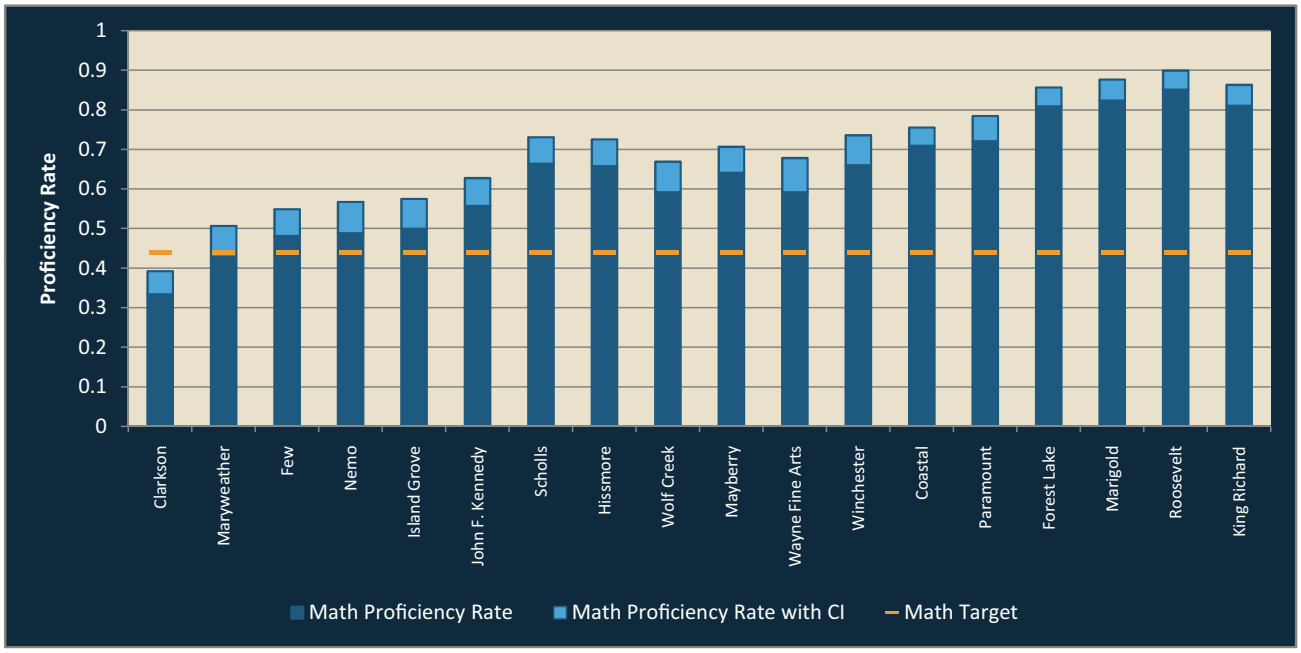

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under New Mexico's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (Maryweather) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

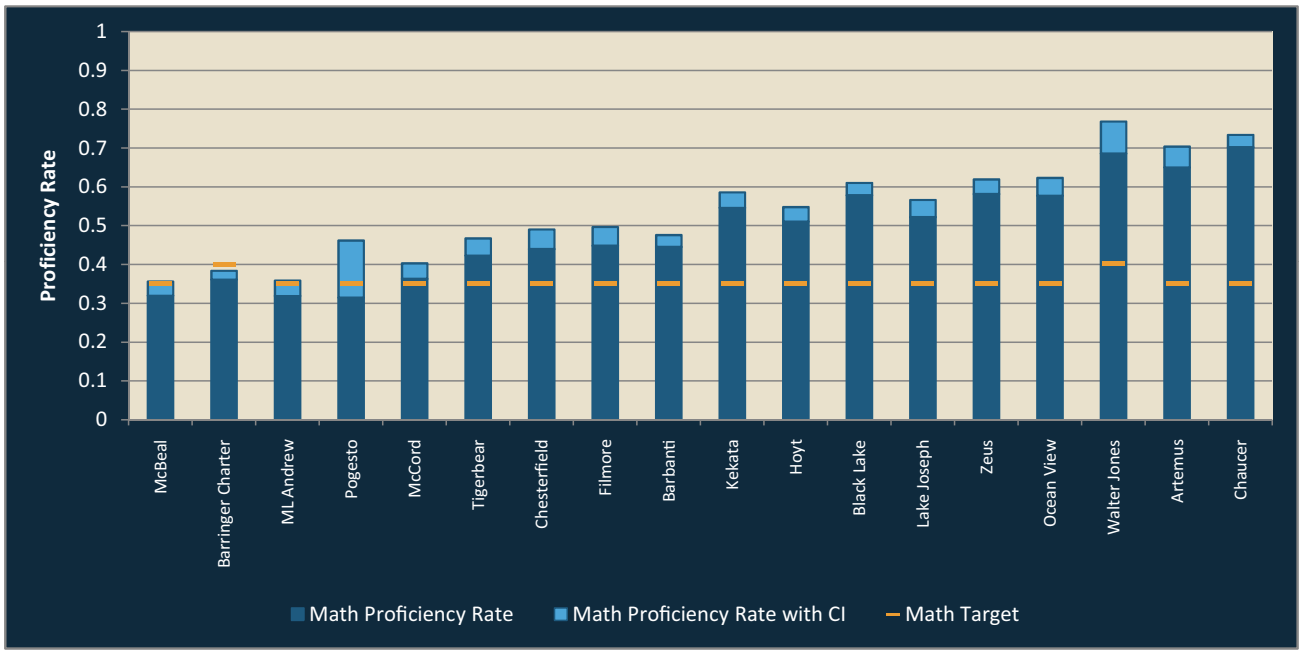

Figure 6. Impact of the confidence interval on middle school math proficiency rates under New Mexico's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that three sample middle schools (McBeal, ML Andrew, and Pogesto) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under New Mexico's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under New Mexico's 2008 AYP rules. **Only 4 of 18 elementary schools (Winchester, Marigold, Roosevelt, and King Richard) made AYP.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. All passing schools are in the right half of the figure, meaning that the highest average performing students were found in these schools.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 New Mexico AYP rules.

Table 2. Elementary school subgroup performance of sample schools under the 2008 New Mexico AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	33.4%	42.3%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Maryweather	42.9%	51.1%	Y	N	N	N	N	N	Y	N	Y	Y			Y	N			Y	Y	14	7	50%	N	1
Few	48.1%	54.3%	Y	Y	N	N	N	N	Y	Y	Y	Y			Y	Y			Y	Y	14	10	71%	N	1
Nemo	48.8%	67.9%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Island Grove	50.0%	67.5%	Y	Y	N	N	N	N	Y	Y					N	Y			Y	Y	12	7	58%	N	4
JFK	55.8%	61.2%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	3
Scholls	66.4%	69.5%	Y	Y	Y	N	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	13	93%	N	7
Hissmore	65.8%	73.3%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	59.2%	67.6%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	5
Alice Mayberry	64.1%	75.4%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	9
Wayne Fine Arts	59.2%	83.3%	Y	Y					Y	Y	N	Y							Y	Y	8	7	88%	N	21
Winchester	66.0%	79.1%	Y	Y	Y	Y			Y	Y					Y	Y			Y	Y	10	10	100%	Y	22
Coastal	70.9%	76.0%	Y	Y	Y	N	Y	N	Y	Y	Y	Y			Y	Y			Y	Y	14	12	86%	N	3
Paramount	72.1%	76.1%	Y	Y	Y	Y	N	N	Y	Y					Y	Y			Y	Y	12	10	83%	N	7
Forest Lake	81.0%	84.9%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	82.4%	87.0%	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y			Y	Y	14	14	100%	Y	10
Roosevelt	85.2%	92.2%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	28
King Richard	81.1%	89.5%	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y			Y	Y	12	12	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Of 18 middle schools in our sample, only 2 made AYP—one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' overall math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates are "increased" by the application of the confidence interval. The orange lines show the AMO needed to meet AYP. These figures show that one of the sample elementary schools (Maryweather) and three middle schools (McBeal, ML Andrew, and Pogesto) are assisted by the confidence intervals. However, of the latter three, only Pogesto also meets all of its subgroup targets in order to make AYP (see Figure 4).

The effect of confidence intervals on schools' proficiency rates in reading is much the same (not shown). In reading, just one elementary school (Few) and one middle school (McBeal) met the overall target with the confidence interval, but we know from Figures 3 and 4 that both schools

Table 3. Middle school subgroup performance of sample schools under the 2008 New Mexico AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	32.0%	52.7%	Y	Y	N	N	N	N	N	N	N	Y	Y	Y	N	N	N	Y	Y	Y	18	8	44%	N	0
Barringer Charter	36.1%	57.1%	N	Y	N	N			N	Y	N	N			Y	Y			Y	Y	12	6	50%	N	0
ML Andrew	31.9%	55.9%	Y	Y	N	N	N	N	N	N	N	N			N	Y			Y	Y	14	5	36%	N	0
Pogesto	31.5%	66.7%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	15
McCord Charter	36.3%	59.2%	Y	Y	N	N	N		N	N	N	N			N	Y			Y	Y	13	5	38%	N	0
Tigerbear	42.3%	56.9%	Y	Y	N	N			Y	Y	N	N							Y	Y	10	6	60%	N	0
Chesterfield	44.0%	58.6%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	1
Filmore	44.9%	67.4%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	1
Barbanti	44.5%	62.8%	Y	Y	N	N	N	N	N	N					Y	Y			Y	Y	12	6	50%	N	0
Kekata	54.6%	66.7%	Y	Y	N	N	N	N	Y	Y	Y	N			Y	N			Y	Y	14	8	57%	N	0
Hoyt	51.1%	69.2%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	2
Black Lake	57.9%	69.2%	Y	Y	N	N	Y	N	Y	Y	Y	N	Y	Y	Y	Y			Y	Y	16	12	75%	N	0
Lake Joseph	52.2%	74.3%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	58.2%	70.5%	Y	Y	N	N	N	N	Y	Y	Y	Y			Y	Y			Y	Y	14	10	71%	N	1
Ocean View	57.7%	80.9%	Y	Y	Y	Y	N	N	Y	Y			Y	Y	Y	Y			Y	Y	14	12	86%	N	2
Walter Jones	68.6%	80.6%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	65.0%	79.2%	Y	Y	Y	N			Y	Y			Y	Y	Y	N			Y	Y	12	10	83%	N	3
Chaucer	70.2%	85.3%	Y	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	14	88%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

still failed to meet targets for some of their subgroups. Overall, **the application of the confidence interval had only modest impact on final AYP decisions for the sample elementary and middle schools in New Mexico.**[9]

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still pass AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's

[9] In the current analyses, confidence intervals were applied to both the overall student population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 4. Summary of subgroup performance of sample elementary schools under 2008 New Mexico AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	8	12
Students with limited English proficiency	10	6	7
Low-income students	18	1	2
African-American students	9	1	1
Asian/Pacific Islander students	1	0	0
Hispanic students	12	2	2
American Indian/Alaska Native students	0	0	0
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under 2008 New Mexico AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	14	14
Students with limited English proficiency	11	9	10
Low-income students	18	5	4
African-American students	11	5	7
Asian/Pacific Islander students	5	0	0
Hispanic students	14	3	3
American Indian/Alaska Native students	1	1	0
White students	18	0	0

minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 New Mexico rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Almost all schools met their reading and math targets for their overall school population.
- Just two elementary schools (Clarkson and Mary-

Table 6. Comparisons between schools that did and didn't make AYP in New Mexico, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	4	14	2	16
Average student body size	225	328	124	951
Average % low income	14	56	42	45
Average % nonwhite	25	45	27	46
Average performance†	7.51	-0.57	0.40	-0.11
Average % growth‡	126	112	109	97
Average number of targets to meet	11	11	7	13

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

weather) failed to meet the reading targets for their overall school population. One failed to meet its math target for the overall population.

- Only one middle school (Barringer) failed to meet its overall math target, and none failed to meet overall reading targets.

- Other subgroups (low income, Hispanic, and African American, among others) performed fairly well at the elementary level.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. First, the performance of SWDs is proving challenging for schools under New Mexico's system, where this subgroup tends to have enough students to meet the state's minimum *n* of 25. In fact, all but one middle school in the study with qualifying SWD subgroups failed to make AYP (Ocean View Middle missed because of its students with LEP subgroup). Students with LEP and African American students are also struggling to meet the state's middle school targets (which are not as problematic for Hispanic or low-income students).

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that New Mexico's NCLB accountability system is, in most respects, behaving like those in other states. For example, Roosevelt, Winchester, and King Richard are among the schools that made AYP in the greatest number of states—28, 22, and 14, respectively. And these schools all made AYP in New Mexico, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in New Mexico.

But New Mexico is also home to a few anomalies. First, consider Wayne Fine Arts (see Table 2). It made AYP in 21 of the 28 states in our sample, yet failed to make AYP in New Mexico. In examining Table 2, we can see that the subgroup of African American students failed to meet its target in math. Second, look at Pogesto Middle School (Table 3). Even with its relatively low average performance, it made AYP in New Mexico, but failed to do so in 13 of 28 states. Like Wayne Fine Arts, its AYP success in New Mexico is most likely attributable to the relatively small number of targets (six) it has to meet, as shown in Figure 4.

This is consistent with the patterns shown in Table 6, which compares schools that do and don't make AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also have much smaller student populations and much lower percentages of nonwhite students. Surprisingly, though, the elementary schools that make AYP have the same number of subgroups (and thus same targets to meet). Middle schools that make AYP have slightly higher performing students, on average, than middle schools that don't, but have drastically smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

This study examined evaluated the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under New Mexico's AYP rules (and AMOs for 2008). Among this sample, only 4 elementary schools and 2 middle schools—6 in all from a total of 36—would have made AYP in New Mexico. Looking across the 28 state accountability systems examined in the study, this puts New Mexico roughly in the middle of the sample distribution, as shown in Figure 1. The fairly high failure rate in New Mexico is perhaps partly explained by the state's minimum *n* size for subgroups, which tends to be smaller than those used in most other states, meaning it holds more subgroups accountable for performance (this despite the state's low overall cut scores in reading and low annual proficiency targets in math and reading).

Because the overriding goal of NCLB is to eliminate education disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, New Mexico's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as SWD, LEP, or African American students. Almost all of the sample schools made AYP in New Mexico for their student populations as a whole. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils aren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if the participation guidelines for English language learners and students with disabilities are more generous under the current state assessment system,[10] doesn't the massive failure of these students (particularly in middle school) to meet New Mexico's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and students with disabilities, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every

[10] See footnote 5.

single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

North Dakota

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 towards these goals.

This report examines North Dakota's NCLB accountability system—particularly how its various rules, criteria and practices result in schools either making AYP—or not making AYP. It also gauges how tough North Dakota's system is compared with other states. For this study, we selected 36 schools from around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether or not each would make AYP under North Dakota's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against North Dakota's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **16 of 18 elementary schools** and **all of the 18 middle schools** in our sample **failed to make adequate yearly progress** in 2008 under North Dakota's accountability system. (This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.)
- Looking across the 28 state accountability systems examined in the study, we find that the number of schools making AYP in North Dakota is exceeded in 20 other sample states (five states tie with North Dakota, each with two elementary schools making AYP). In addition, North Dakota is one of five states with *zero* passing middle schools in the sample (see Figure 1).
- Many of the schools in our sample that failed to make AYP in North Dakota are meeting expected targets for their overall populations but failing because of the performance of individual subgroups, particularly students with disabilities and English language learners.[2]
- **Two sample schools failed to make AYP in North Dakota that made AYP in most other states. This is likely due to the fact that North Dakota's minimum subgroup size of 10 is small, compared to other states in the study.[3] In addition, North Dakota's annual targets for proficiency are relatively ambitious.**

Only two of the 36 schools in our sample make AYP in 2008 under **North Dakota's** accountability system. The greatest contributing factor to the high failure rate is that North Dakota's minimum subgroup size is 10, which is considerably smaller than most other states we examined. This means that schools in North Dakota will have more accountable subgroups than would similar schools in other states. On the other hand, North Dakota's proficiency standards are about average when compared to the other states in the study. The state also uses a 99 percent confidence interval which provides schools with greater leniency than the more commonly used 95 percent confidence interval. The latter likely explains why two sample schools were able to make AYP in North Dakota.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the North Dakota State Assessment (NDSA).

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup

[3] The state of North Dakota does not have a minimum school size, so it has a large number of very small schools. In addition, the state's population has been declining in recent years. The U.S. Census Bureau (2002) lists North Dakota's population at a little over 642,000, 47th in the United States. Therefore, smaller subgroup sizes are likely warranted.

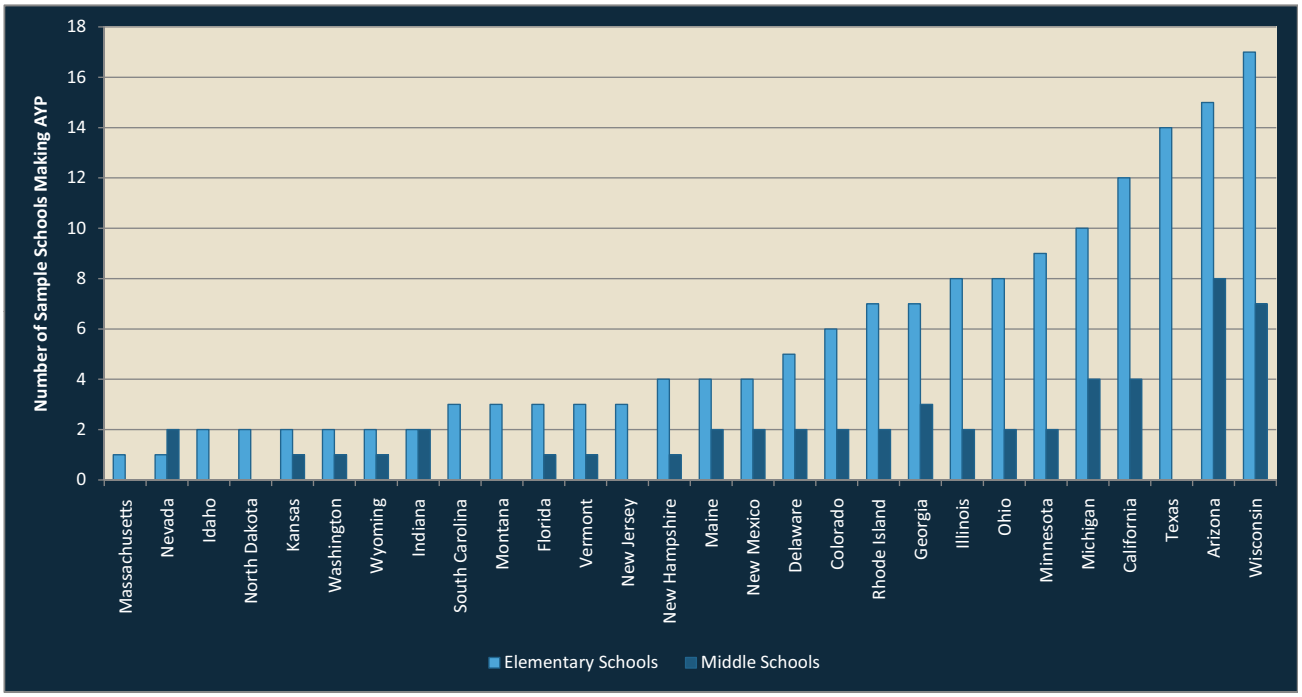

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- As in other states, middle schools have greater difficulty reaching AYP in North Dakota than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower.

- Part of the reason all middle schools failed to make AYP in North Dakota is that its schools have enough low-income, disabled, or limited English proficiency (LEP)[4] students to qualify as separate subgroups. Each of our sample middle schools in North Dakota has one or more of these subgroups and each failed to make AYP. Likewise, many elementary schools with enough students qualifying for these subgroups also failed, though they tended to reach their math targets more often than their reading targets.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance North Dakota's tests and those of 25 other states to the Northwest Evaluation Association's Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency stan-

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

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 North Dakota State Assessment (NDSA), the standardized state test. 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. Our 2005–2006 MAP data do not capture these subgroup nuances. In this study, however, no valid MAP scores were omitted from consideration.

dards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state. This means that a school making AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005-06 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the North Dakota State Assessment (NDSA) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that North Dakota's definitions of proficiency generally ranked about average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the North Dakota test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under North Dakota AYP rules for 2008. In other words, the school data and our proficiency cut score es-

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

[7] We gave all schools in our sample pseudonyms in this report.

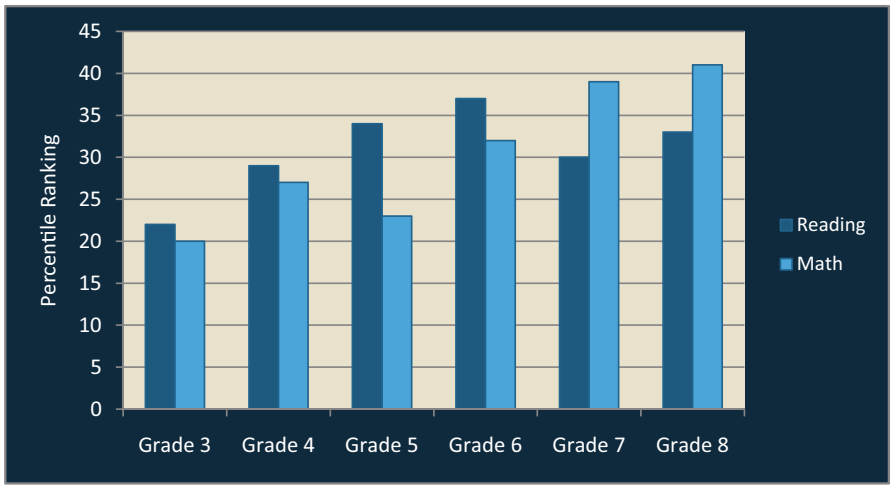

Figure 2. North Dakota reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of North Dakota's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of North Dakota's cut scores are below the 45th percentile.

timates are from academic year 2005–2006, but we are applying them against North Dakota's 2008 AYP rules.

Table 1 shows the pertinent North Dakota AYP rules that were applied to elementary and middle schools in this study. North Dakota's minimum subgroup size is 10, which is considerably smaller than most other states we examined.[8] This means that schools in North Dakota will have more accountable subgroups than would similar schools in other states. North Dakota's annual targets also differ by grade and subject. For example, 66.7% of grade 8 math students are expected to be proficient in 2008; the percentage for grade 3 reading students is 82.6%.

Most states examined also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. However, North Dakota's 99% confidence interval provides schools with greater leniency than the more commonly used 95% confidence interval. So, for instance, while schools are supposed to get 82.6% of their students in grade 3 to the "proficient" level on the state reading test, and 82.6% of the students in each subgroup, applying the confidence interval means that the real target can be lower.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each school's subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population

[8] The state of North Dakota does not have a minimum school size, so it has a large number of very small schools. In addition, the state's population has been declining in recent years. The U.S. Census Bureau (2002) lists North Dakota's population at a little over 642,000, 47th in the United States. Therefore, smaller subgroup sizes are likely warranted.

Table 1. North Dakota AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 10	
	SWDs: 10	
	Low-income students: 10	
	LEP students: 10	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	82.6
Grade 4	65.1	82.6
Grade 5	n/a	82.6
Grade 6	n/a	80.7
Grade 7	n/a	80.7
Grade 8	61.4	80.7
MATH		
Grade 3	n/a	72.9
Grade 4	45.7	72.9
Grade 5	n/a	72.9
Grade 6	n/a	66.7
Grade 7	n/a	66.7
Grade 8	33.3	66.7

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not applicable

and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under North Dakota's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under North Dakota's 2008 AYP rules. **Only 2 elementary schools (Wayne Fine Arts and Roosevelt) made AYP while 16 failed to make it**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The two schools making AYP are in the right half of the figure, meaning that they are among the schools that contain the higher average performing students. Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 North Dakota AYP rules. Not a single middle school in the sample makes AYP under the North Dakota rules.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to

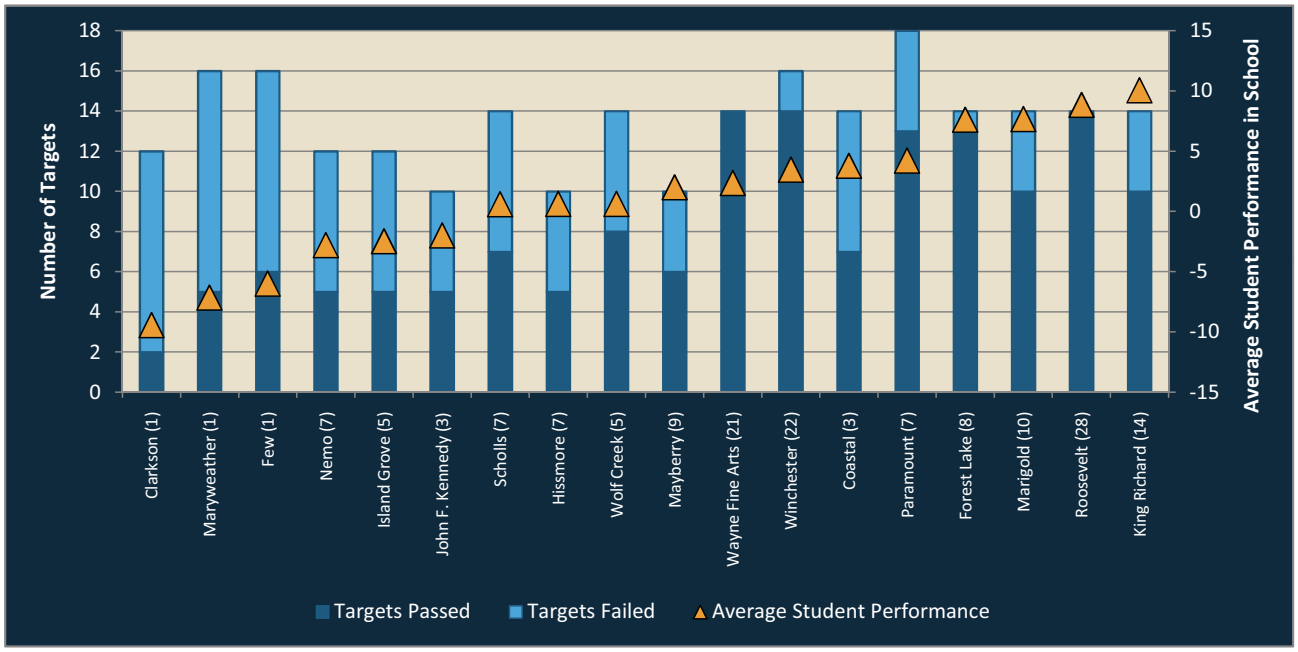

Figure 3. AYP performance of the elementary school sample under North Dakota 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under North Dakota's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Marigold Elementary, for example, met ten of its fourteen targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school makes AYP in the study.

which these proficiency rates are increased by the application of the confidence interval. The orange lines show the annual measurable objective (or annual target) needed to meet AYP. The figures show that only one of the sample elementary schools (Maryweather) and three of the middle schools (Tigerbear, Chesterfield, and Filmore) were assisted by the confidence intervals (note how the orange lines fall within the light blue bands). However, we know that all of these schools still failed to make AYP because of low subgroup performance (see Figures 3 and 4). Tigerbear, for instance, didn't meet nine of its twelve targets.

The effect of confidence intervals on the reading proficiency rates for elementary and middle schools shows largely the same pattern (not shown). In reading, two elementary schools (Mayberry and Paramount) and one middle school (Pogesto) met the overall target with the confidence interval, but we know from Figures 3 and 4 that these schools still fail to meet targets for subgroups. In short, **the application of the confidence interval had little or no impact on whether schools achieved their overall math and reading targets in North Dakota (or whether they made AYP).**[9]

Where Do Schools fail?

Figures 3 and 4 illustrate the number of subgroup targets at each of the sample schools, but these figures do not indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's complexity and length.

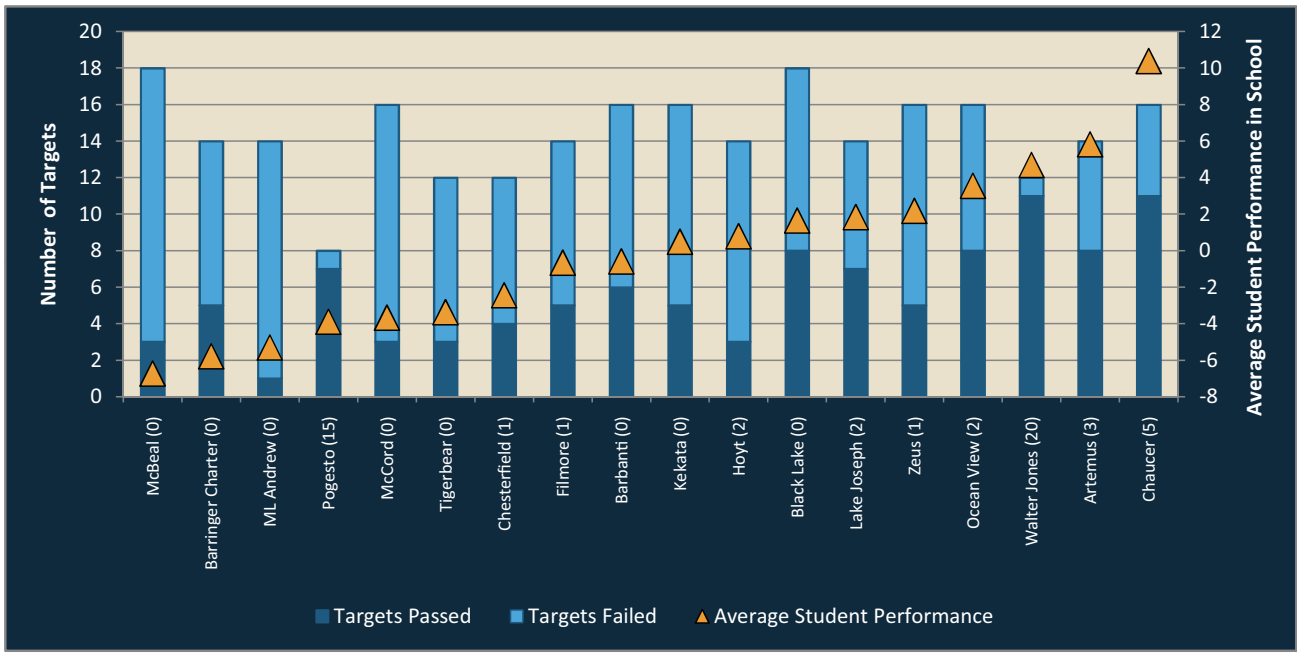

Figure 4. AYP performance of the middle school sample under North Dakota 2008 AYP rules

Note: This figure indicates how each of the middle schools within the sample fared under North Dakota's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Pogesto, for example, met seven of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school would make AYP in the study.

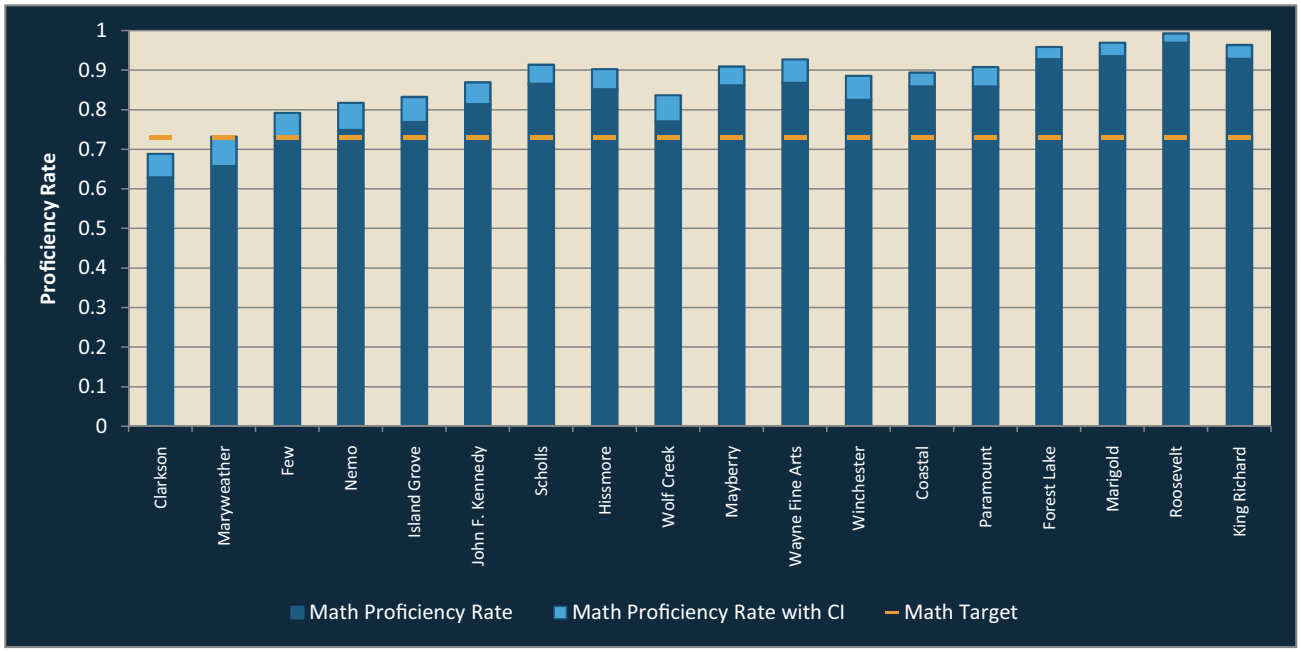

Figure 5. Impact of the confidence Interval on elementary school math proficiency rates under North Dakota 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (Maryweather) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

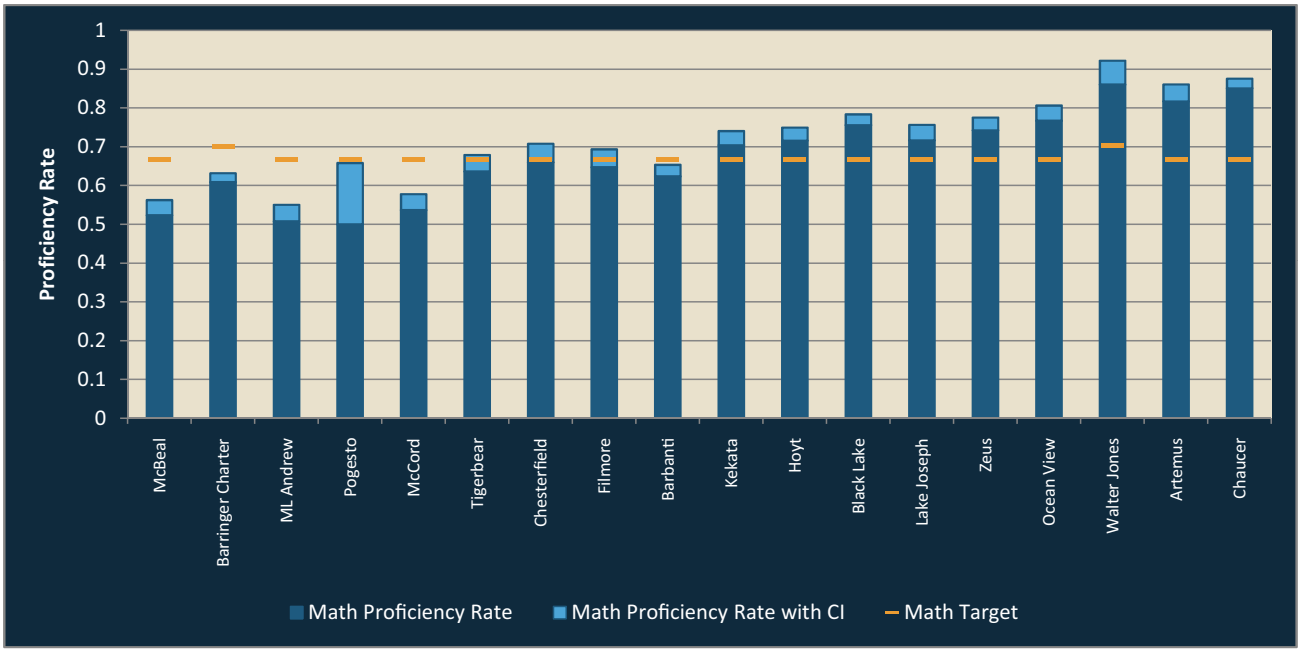

Figure 6. Impact of the confidence interval on middle school math proficiency rates under North Dakota 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that three of the sample middle schools (Tigerbear, Chesterfield, and Filmore) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. While all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP purposes include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American (AA), Asian/Pacific Islander (Asian), Hispanic/Latino (Hispanic), American Indian/Alaska Native (AI/AN), and White. Tables 2 and 3 also show whether a school made AYP under the North Dakota rules, and the total number of states within the study in which that school makes AYP.

The school-by-school findings in Tables 2 and 3 are summarized as shown:

- One of the 18 elementary schools (Clarkson) failed to meet both reading and math targets for its overall student population, while nine others (Maryweather, Few, Nemo, Island Grove, JFK, Scholls, Hissmore, Wolf Creek, and Coastal) failed to meet their overall reading targets.
- Twelve middle schools (McBeal, Barringer, ML Andrew, McCord, Tigerbear, Chesterfield, Filmore, Barbanti, Kekata, Hoyt, Black Lake, and Zeus) failed to meet their overall reading targets, and six (McBeal, Barringer, ML Andrew, Pogesto, McCord, and Barbanti) failed in math.
- One elementary school (Forest Lake) met every target except for its reading target for students with disabilities.
- One middle school (Walter Jones) met all targets for every subgroup except for its African American population.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively.[10] The performance of SWDs is proving challenging

[10] Recall that elementary students generally do better on North Dakota's math test than middle school students, partly because North Dakota's cut scores are lower in math than in reading at the elementary level (see Figure 2).

Table 2. Elementary subgroup performance of sample schools under the 2008 North Dakota AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	62.9%	43.1%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	1
Maryweather	65.8%	52.5%	Y	N	N	N	N	N	N	N	Y	N			N	N	Y	N	Y	Y	16	5	31%	N	1
Few	73.2%	55.3%	Y	N	N	N	N	N	Y	N	Y	N			Y	N	Y	N	Y	N	16	6	38%	N	1
Nemo	74.9%	71.6%	Y	N	N	N			Y	N	N	N			Y	N			Y	Y	12	5	42%	N	7
Island Grove	76.9%	68.3%	Y	N	N	N	N	N	Y	N					Y	N			Y	Y	12	5	42%	N	4
JFK	81.4%	66.0%	Y	N	Y	N			Y	N	Y	N							Y	N	10	5	50%	N	3
Scholls	86.6%	71.7%	Y	N	N	N	Y	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	7
Hissmore	85.2%	74.8%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	7
Wolf Creek	77.1%	73.1%	Y	N	N	N	Y	N	Y	N			Y	Y	Y	N			Y	Y	14	8	57%	N	5
Alice Mayberry	86.2%	78.9%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	9
Wayne Fine Arts	86.8%	83.3%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	21
Winchester	82.5%	82.0%	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	14	88%	N	22
Coastal	85.9%	75.2%	Y	N	N	N	Y	N	Y	N	Y	N			Y	N			Y	Y	14	7	50%	N	3
Paramount	85.9%	78.7%	Y	Y	Y	N	N	N	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	18	13	72%	N	7
Forest Lake	92.8%	86.6%	Y	Y	Y	N			Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	14	13	93%	N	8
Marigold	93.5%	84.8%	Y	Y	Y	N	Y	N	Y	N			Y	Y	Y	N			Y	Y	14	10	71%	N	10
Roosevelt	97.0%	92.9%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	28
King Richard	92.9%	87.8%	Y	Y	Y	N	Y	N	Y	N			Y	Y	Y	N			Y	Y	14	10	71%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

for schools under North Dakota's system where this subgroup tends to have enough students to meet the state's minimum *n* of 10. In fact, all but two elementary schools and all but one middle school in the study with qualifying SWDs reading subgroups failed to make AYP. Students with LEP and low-income students are also struggling to meet the state's targets; almost every single school with a large enough population to qualify as a separate subgroup failed to meet its targets for these students (though they tend to do better in math at the elementary level).

Other state reports contain a section comparing some of the characteristics of the sample schools that made AYP versus those that did not. In North Dakota, none of the sample middle schools made AYP, and among elementary schools, there were no striking differences among schools that did and didn't make AYP. The one exception (rather expected) was that schools that made AYP had students with higher average performance than did schools that didn't make it, as measured by NWEA reading and math tests.

Table 3. Middle school subgroup performance of sample schools under the 2008 North Dakota AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	52.4%	54.2%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	N	18	3	17%	N	0
Barringer Charter	60.9%	60.2%	N	N	N	N	Y	Y	N	N	N	N			Y	N			Y	Y	14	5	36%	N	0
ML Andrew	50.8%	59.8%	N	N	N	N	N	N	N	N	N	N			N	N			Y	N	14	1	7%	N	0
Pogesto	50.0%	68.5%	N	Y					Y	Y					Y	Y			Y	Y	8	7	88%	N	15
McCord Charter	53.7%	61.4%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N			Y	N	16	3	19%	N	0
Tigerbear	63.6%	59.6%	Y	N	N	N			N	N	N	N			Y	N			Y	N	12	3	25%	N	0
Chesterfield	66.0%	60.9%	Y	N	N	N			N	N	N	N			Y	Y			Y	N	12	4	33%	N	1
Filmore	64.7%	69.9%	Y	N	N	N	N	N	N	N			Y	Y	N	N			Y	Y	14	5	36%	N	1
Barbanti	62.4%	64.8%	N	N	N	N	N	N	N	N	Y	Y	Y	Y	N	N			Y	Y	16	6	38%	N	0
Kekata	70.4%	68.5%	Y	N	N	N	N	N	N	N	N	N	Y	Y	N	N			Y	Y	16	5	31%	N	0
Hoyt	71.6%	71.7%	Y	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	3	21%	N	2
Black Lake	75.6%	71.2%	Y	N	N	N	N	N	Y	N	N	N	Y	Y	Y	N	Y	Y	Y	N	18	8	44%	N	0
Lake Joseph	71.7%	77.2%	Y	Y	N	N	N	N	Y	N	Y	Y			N	N			Y	Y	14	7	50%	N	2
Zeus	74.2%	72.7%	Y	N	N	N	N	N	N	N	Y	N	Y	Y	N	N			Y	N	16	5	31%	N	1
Ocean View	76.7%	83.2%	Y	Y	N	N	N	N	N	N	Y	Y	Y	Y	N	N			Y	Y	16	8	50%	N	2
Walter Jones	86.0%	83.4%	Y	Y	Y	Y			Y	Y	Y	N			Y	Y			Y	Y	12	11	92%	N	20
Artemus	81.7%	80.5%	Y	Y	N	N			N	N	Y	Y	Y	Y	N	N			Y	Y	14	8	57%	N	3
Chaucer	85.1%	86.9%	Y	Y	N	N	N	N	Y	N	Y	Y	Y	Y	Y	Y			Y	Y	16	11	69%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that North Dakota's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt and Wayne Fine Arts each made AYP in many states—28 and 21, respectively. And these schools made AYP in North Dakota, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed AYP in North Dakota.

But North Dakota is also home to a few anomalies. First, consider Winchester Elementary (see Figure 3). It made AYP in 22 of the 28 states in our sample, yet not in North Dakota. Examining Table 2, one can see that Winchester didn't meet reading targets for its LEP or SWD subgroups, although the school's overall reading proficiency rate was 82%. Second, look at Walter Jones Middle School (Figure 4). Even with its relatively high average performance it didn't make AYP in North Dakota, but made AYP in 20 of 28 states. Like Winchester, it missed the AYP mark in North Dakota probably because of North Dakota's relatively small minimum *n*.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 North Dakota AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	18	10	16
Students with limited English proficiency	13	5	11
Low-income students	18	2	14
African-American students	13	1	8
Asian/Pacific Islander students	6	0	0
Hispanic students	15	2	11
American Indian/Alaska Native students	3	0	2
White students	18	0	2

Table 5. Summary of subgroup performance of sample middle schools under the 2008 North Dakota AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	17	16	16
Students with limited English proficiency	13	12	12
Low-income students	18	13	16
African-American students	16	9	11
Asian/Pacific Islander students	10	0	0
Hispanic students	18	11	14
American Indian/Alaska Native students	2	1	1
White students	18	0	7

Note: The relatively high number of qualifying subgroups for African American, Asian/Pacific Islander, and Hispanic students is largely due to Nebraska's minimum n size of 10.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under North Dakota's AYP rules (and AMOs) for 2008. Among this sample, only two elementary schools and no middle schools—two in all from a sample of 36—would have made AYP in North Dakota. Looking across the 28 state accountability systems examined in the study, this puts North Dakota at the low end of the distribution in terms of the number of schools making AYP (see Figure 1). North Dakota's small minimum *n* size and fairly high AMOs likely lead to the large number of schools that failed to make AYP.

The overriding goal of the federal NCLB is to eliminate education disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, North Dakota's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income students. Many of the sample schools met the North Dakota reading and math targets for their student populations as a whole (more so in math than reading). In the pre-NCLB era, such schools might have been considered to be effective or at least not needing improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, etc. has made those students visible. That is surely a good thing.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? And in the case of North Dakota, that small subgroup sizes and high annual targets make it nearly impossible for schools to be viewed as successful? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[11] doesn't the failure of these students to meet North Dakota's targets (especially at the middle school level) indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

[11] See footnote 5.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Ohio

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Ohio's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP or not making AYP. It also gauges how tough Ohio's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Ohio's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Ohio's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **10 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Ohio's accountability system. (This rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students.)
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools that made AYP in Ohio was exceeded in just 6 other sample states** (Ohio and Illinois tie with 8 elementary schools making AYP) (see Figure 1).[2]
- Nearly all of the schools in our sample that failed to make AYP in Ohio are meeting expected targets for their overall populations[3] but failing because of the performance of individual subgroups, particularly students with disabilities (SWDs)[4] and English language learners.
- A few sample schools that made AYP in Ohio failed to make AYP in most other states. **This is most likely because Ohio's proficiency standards are relatively easy compared to other states, and Ohio's minimum *n* (number of students in sample) size for SWDs is higher than other states, meaning that**

Ohio falls in the upper end of the state distribution in terms of the number of schools that make AYP. In fact, a few sample schools make AYP in Ohio that fail to make AYP in most other states. This is likely because Ohio's proficiency standards are relatively easy compared to other states (most of Ohio's cut scores are below the 35th percentile). Additionally, while Ohio's minimum *n* size for most of its subgroups is a little lower than in other states (30), the state raises its subgroup size to 45 for students with disabilities, meaning fewer of these students are held separately accountable than in other jurisdictions. On the other hand, Ohio does not apply confidence intervals (or margins of error) to their measurements of student proficiency rates. This means that schools in Ohio will have a more difficult time meeting their targets than schools in states that do use confidence intervals.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Ohio Achievement Test.

[2] In 2006, Ohio received approval from the U.S. Department of Education to use a student growth model in its state accountability plan. The data in this study are drawn from 2005–2006 and do not reflect student growth calculations in any way.

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup.

[4] SWDs are defined as those students following individualized education plans.

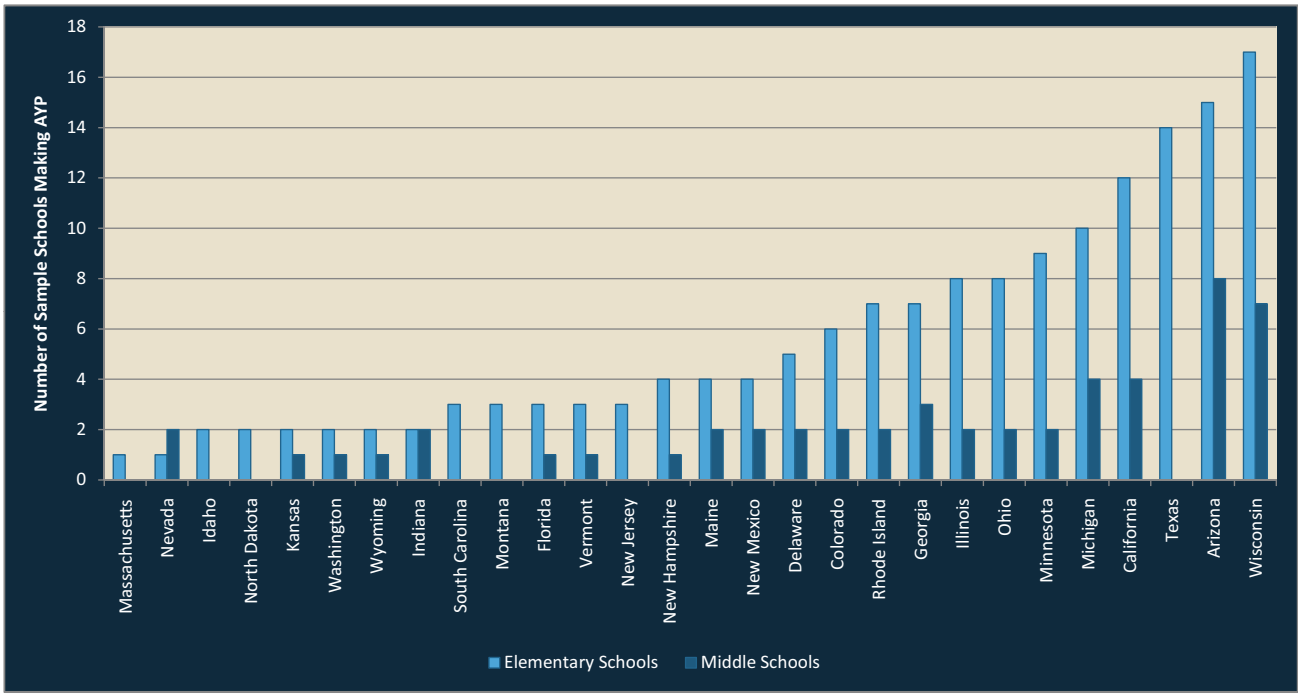

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

fewer SWD subgroups in Ohio (especially at the elementary level) are likely to be held separately accountable for performance.

- As in other states, schools with fewer subgroups attained AYP more easily in Ohio than schools with more subgroups, even when their average student performance is lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As in other states, middle schools in Ohio had greater difficulty reaching AYP than did elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of whether or not a school will make AYP under Ohio's system is whether it has enough limited English proficient (LEP) students[5] to qualify as a separate subgroup. Almost every single school with even one such subgroup failed to make AYP, in part because these students did not meet the state's targets in reading and math.[6]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Ohio's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child

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

[6] We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Ohio Achievement Test, the standardized state test. 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.

Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[7] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than 10 pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[8]

Proficiency cut score estimates for the Ohio Achievement Test (OAT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Ohio's definitions of proficiency generally ranked below average compared with the standards set by the other 25 states in that study.

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

[8] We gave all schools in our sample pseudonyms in this report.

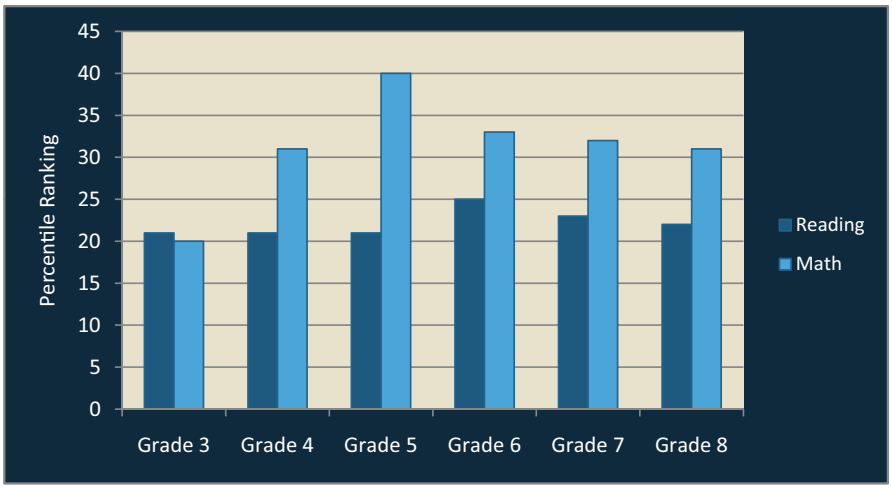

Figure 2. Ohio reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Ohio's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Ohio's cut scores are at or below the 40th percentile.

Table 1. Ohio AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 45	
	Low-income students: 30	
	LEP students: 30	
CI	**Applied to proficiency rate calculations?**	
	Not used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	77.0
Grade 4	36.0	74.6
Grade 5	n/a	74.6
Grade 6	n/a	80.6
Grade 7	n/a	74.9
Grade 8	n/a	79.0
MATH		
Grade 3	n/a	68.5
Grade 4	36.0	73.7
Grade 5	n/a	59.7
Grade 6	n/a	64.1
Grade 7	n/a	57.8
Grade 8	n/a	58.0

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not applicable

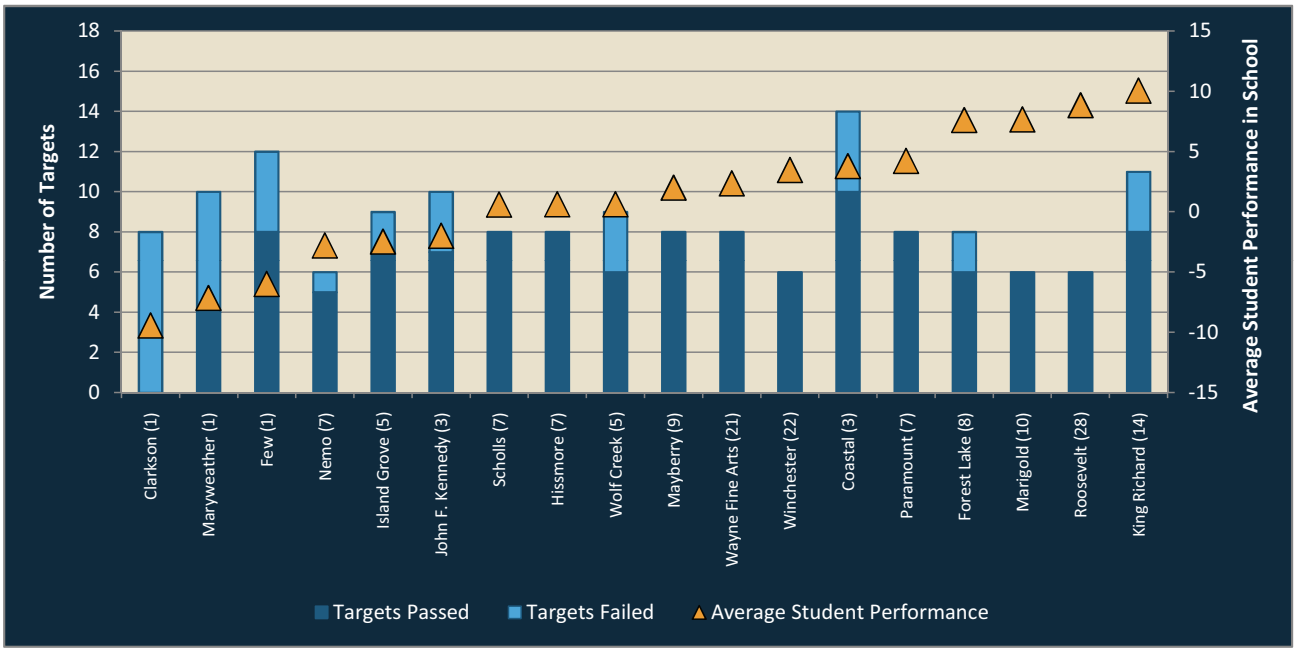

Figure 3. AYP performance of the elementary school sample under Ohio 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Ohio's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Forest Lake, for example, met 6 of its 8 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

These cut scores were used to estimate whether students would have scored as proficient or better on the Ohio test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Ohio AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Ohio's 2008 AYP rules.

Table 1 shows the pertinent Ohio AYP rules that we applied to elementary and middle schools in the current study. Ohio's minimum subgroup size is 30 for three of the four reporting groups (race/ethnicity, low income, and English proficiency), but 45 for the fourth group (students with disabilities), which is higher than most other states we examined.[9]

Specifically, most states have a subgroup size of around 35-40 for reporting purposes but typically don't alter *n* sizes based on particular subgroups. Also unlike most other states, Ohio does not apply confidence intervals (or margins of statistical error) to its measurements of student proficiency rates. This means that schools in Ohio will have a more difficult time meeting their targets than schools in states that do use confidence intervals. Annual targets in Ohio also differ by grade and subject matter (e.g., 57.8% of seventh graders are expected to be proficient in math in 2008; that number changes to 80.6% for sixth graders in reading).

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient stu-

[9] School size and *n* size, however, are related (e.g., it makes sense for small schools to have small *n* sizes).

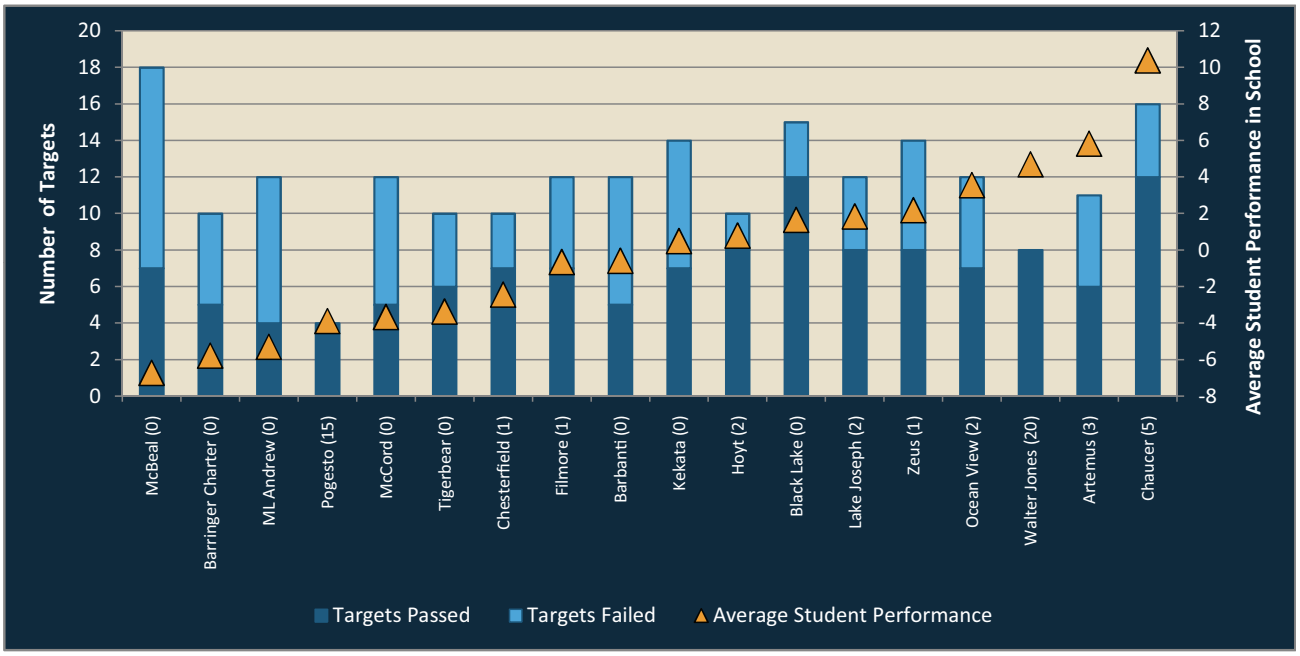

Figure 4. AYP performance of the middle school sample under Ohio 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under Ohio's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup did not make AYP, so any light blue means that the school failed. Hoyt, for example, met 8 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school, and its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

dents within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Ohio's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Ohio's 2008 AYP rules. **Eight elementary schools made AYP (Scholls, Hissmore, Mayberry, Wayne Fine Arts, Winchester, Paramount, Marigold, and Roosevelt) and 10 failed to make AYP.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The majority of the schools that made AYP are in the right half of the figure, meaning that higher performing students were found at these schools.

Table 2. Elementary subgroup performance of sample schools under the 2008 Ohio AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	54.5%	55.5%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	59.8%	63.9%	N	Y			N	N	N	N					N	Y			Y	Y	10	4	40%	N	1
Few	66.0%	67.0%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	1
Nemo	69.8%	78.1%	Y	Y					N	Y									Y	Y	6	5	83%	N	7
Island Grove	71.4%	77.0%	Y	Y				N	Y	Y					N	Y			Y	Y	9	7	78%	N	4
JFK	74.0%	74.6%	Y	Y	N	N			Y	Y	N	Y							Y	Y	10	7	70%	N	3
Scholls	82.5%	80.3%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Hissmore	81.4%	82.4%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Wolf Creek	73.9%	78.1%	Y	Y				N	N	Y					N	Y			Y	Y	9	6	67%	N	5
Alice Mayberry	80.3%	84.1%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	82.2%	90.2%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	78.3%	86.7%	Y	Y											Y	Y			Y	Y	6	6	100%	Y	22
Coastal	81.8%	82.6%	Y	Y	N	N	N	N	Y	Y	Y	Y			Y	Y			Y	Y	14	10	71%	N	3
Paramount	82.2%	82.1%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	7
Forest Lake	89.8%	90.6%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	91.7%	91.7%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	10
Roosevelt	93.6%	96.9%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	89.5%	94.2%	Y	Y	N		N	Y	Y	Y					N	Y			Y	Y	11	8	73%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Yet almost without regard to average student performance, the schools that made AYP were primarily those with relatively few qualifying subgroups—and thus the fewest targets to meet. For example, Winchester made it, but had only six targets (two targets in reading and math for its overall student population, two more for its Hispanic subgroup, and two more for its white subgroup).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Ohio AYP rules. **Of 18 middle schools in our sample, only 2 made AYP**—one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which have relatively few qualifying subgroups.

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Table 3. Middle school subgroup performance of sample schools under the 2008 Ohio AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	59.7%	65.5%	N	Y	N	N	N	N	N	N	N	Y	Y	Y	N	N	N	Y	Y	Y	18	7	39%	N	0
Barringer Charter	60.0%	71.8%	N	Y	N	N			N	Y	N	Y			Y	Y					10	5	50%	N	0
ML Andrew	58.6%	71.6%	N	Y	N	N			N	N	N	N			N	Y			Y	Y	12	4	33%	N	0
Pogesto	64.8%	75.9%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	60.4%	73.2%	Y	Y	N	N			N	N	N	N			N	Y			Y	Y	12	5	42%	N	0
Tigerbear	68.5%	68.9%	Y	Y	N	N			Y	Y	N	N							Y	Y	10	6	60%	N	0
Chesterfield	73.8%	74.0%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	1
Filmore	70.5%	80.0%	Y	Y	N	N	N	N	Y	Y					N	Y			Y	Y	12	7	58%	N	1
Barbanti	67.7%	75.6%	Y	Y	N	N	N	N	N	N					N	Y			Y	Y	12	5	42%	N	0
Kekata	75.6%	76.7%	Y	Y	N	N	N	N	Y	Y	N	Y			N	N			Y	Y	14	7	50%	N	0
Hoyt	78.2%	80.9%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	2
Black Lake	80.9%	80.9%	Y	Y	N	N	N		Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	15	12	80%	N	0
Lake Joseph	77.3%	84.6%	Y	Y	N	N	N	N	Y	Y					Y	Y			Y	Y	12	8	67%	N	2
Zeus	80.6%	81.6%	Y	Y	N	N	N	N	Y	Y	Y	Y			N	N			Y	Y	14	8	57%	N	1
Ocean View	82.3%	89.4%	Y	Y	N	Y	N	N	N	Y					N	Y			Y	Y	12	7	58%	N	2
Walter Jones	84.3%	89.1%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	83.8%	86.1%	Y	Y		N			N	N			Y	Y	N	N			Y	Y	11	6	55%	N	3
Chaucer	89.5%	92.6%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Ohio rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Overwhelmingly, schools met their targets for their overall student populations. Only one elementary school (Clarkson) failed to meet its math and reading targets for its overall school population. One additional elementary school (Maryweather) failed to meet its overall math target.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Ohio AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	5	5	4
Students with limited English proficiency	7	5	6
Low-income students	17	4	2
African-American students	6	1	0
Asian/Pacific Islander students	0	0	0
Hispanic students	9	5	1
American Indian/Alaska Native students	0	0	0
White students	17	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Ohio AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	15	15	15
Students with limited English proficiency	9	9	8
Low-income students	17	7	5
African-American students	11	6	4
Asian/Pacific Islander students	4	0	0
Hispanic students	14	9	4
American Indian/Alaska Native students	1	1	0
White students	17	0	0

- Three sample middle schools (McBeal, Barringer, and ML Andrew) failed to meet their math targets for their overall populations.
- One elementary school (Nemo) met its math and reading targets for every subgroup except low-income students.
- One elementary school (Forest Lake) met all its targets except for students with disabilities.
- Low-income students tended to meet their annual targets, especially in reading at the elementary level. But all schools with qualifying LEP and SWD subgroups failed to make AYP.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. First, the performance of students with disabilities is proving quite challenging for schools under

Table 6. Comparisons between schools that did and didn't make AYP in Ohio, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	8	10	2	16
Average student body size	256	344	124	951
Average % low income	37	54	42	45
Average % nonwhite	36	45	27	46
Average performance†	3.72	-0.77	0.40	-0.11
Average % growth‡	113	116	109	97
Average number of targets to meet	7	10	6	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

Ohio's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* size of 45. In fact, all but one SWD subgroup in the study (at Ocean View) failed to meets its AYP targets. Students with limited English proficiency are also struggling to meet the state's targets; almost every school with a large enough LEP population to qualify as a separate subgroup failed to meet its reading targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Ohio's NCLB accountability system is, in many respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Ohio, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Ohio.

But Ohio is also home to a few anomalies. First, consider Mayberry Elementary (see Figure 3). It failed to make AYP in 19 of the 28 states in our sample, yet made AYP in Ohio. In examining Table 2, we can see that Mayberry didn't meet the minimum numbers for the LEP or SWD subgroups, which created difficulty for so many other schools within the sample. With fewer accountable subgroups and with relatively easy proficiency standards (Figure 2), Mayberry attained AYP in Ohio, even when other schools with higher average performance failed. This seems to be the case for a few other elementary schools (Hissmore, Paramount, and Marigold) and for at least one middle school (Pogesto).

This is consistent with the patterns shown in Table 6, which compares schools making and not making AYP on a number of academic and demographic dimensions. Within the sample, passing schools do indeed show higher average student performance, but they also differ in the following ways: they have smaller student populations (dramatically so at the middle school level) and fewer subgroups (and thus fewer targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Ohio's AYP rules (and AMOs) for 2008. We found that 8 elementary schools and 2 middle schools—10 in all, from a sample of 36—would have made AYP in Ohio. Looking across the 28 state accountability systems examined in the study, this puts Ohio towards the high end of the sample distribution in terms of the number of schools making AYP (see Figure 1). Part of the reason that some schools made AYP in Ohio and not in other states is that Ohio's proficiency standards are relatively easy. In addition, Ohio's minimum *n* size for SWDs is higher than in other states, meaning that fewer SWD subgroups in Ohio (particularly at the elementary level) are likely to be held accountable for performance.

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Ohio's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income students. Almost all of the sample schools met the Ohio reading and math targets for their overall populations, i.e., without considering subgroup results. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP (and in Ohio, that those subgroup *n* sizes change based on subgroup classification)? Even if actual participation guidelines for English language learners and students with disabilities are more generous under the current state assessment system,[10] doesn't the massive failure of these students, especially in middle schools, to meet Ohio's targets indicate that a new approach is needed for holding schools accountable for their performance? Yes, schools should redouble their efforts to boost achievement for LEP students and students with disabilities, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions,

[10] See footnote 6.

which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Rhode Island

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Rhode Island's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP—or not making AYP. It also gauges how tough Rhode Island's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Rhode Island's system as well as under the systems of twenty-seven other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Rhode Island's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **11 of 18 elementary schools** and **16 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Rhode Island's accountability system. (This failure rate is partly explained by our sample, which intentionally includes some schools with relatively large populations of low-performing students.)
- Looking across the 28 state **accountability systems examined in the study, we find that the number of elementary schools making AYP in Rhode Island was exceeded in just 8 other sample states** (Rhode Island ties Georgia with 7 elementary schools making AYP). (See Figure 1.)
- Many of the schools in our sample that failed to make AYP in Rhode Island are meeting expected targets for their overall populations but didn't make AYP because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[2]
- Two sample schools that failed to make AYP in most other states made AYP in Rhode Island. **This is probably because Rhode Island's minimum subgroup**

Unlike most states, **Rhode Island** measures its student performance with a proficiency index, which gives partial credit for students achieving "partial proficiency." In the short term, the index makes it easier for Rhode Island schools to meet their targets. However, the effect of the index diminishes as the annual targets gradually approach the 100 percent proficiency requirement dictated under NCLB for 2014. Two sample schools make AYP in Rhode Island that fail to make AYP in most other states. This is likely because Rhode Island's minimum subgroup size (45) is larger than in most other states, meaning that schools have fewer accountable subgroups under Rhode Island rules than they would in another state. In addition, Rhode Island's proficiency standards are average when compared to other states, but its annual targets are fairly rigorous. In grades 3-5 reading, for example, Rhode Island requires roughly 84 percent of all subgroups to reach proficiency in order for a school to make AYP in 2008.

[1] A cut score is the minimum score a student must receive on the New England Common Assessment Program in order to be considered proficient under Rhode Island's accountability system.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup.

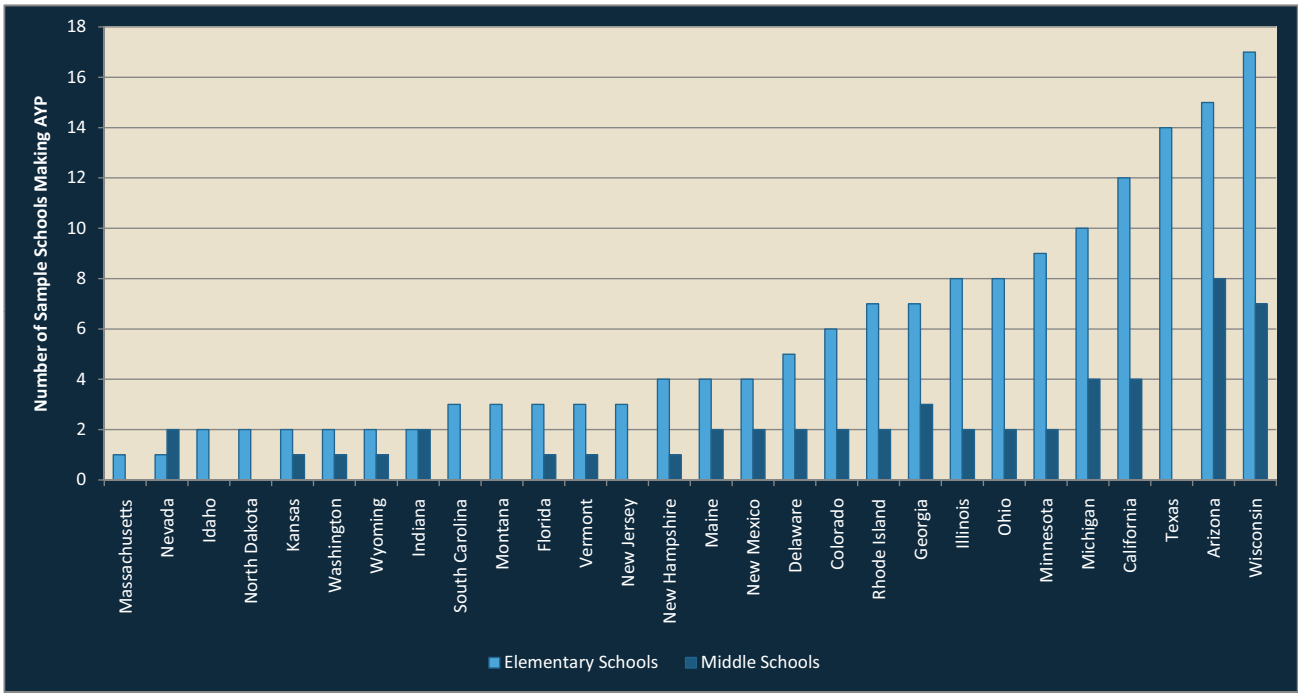

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

sizes are a bit larger than in most other states, meaning that these schools had fewer accountable subgroups under Rhode Island rules.

- Rhode Island's proficiency standards (or cut scores) are above average compared to other states; similarly, the state's annual targets for proficiency are fairly ambitious, particularly at the elementary school level. However, Rhode Island uses a proficiency index, which gives partial credit for students achieving "partial proficiency." **In the short term, the index makes it easier for schools in Rhode Island to meet their targets, although the effect of the index diminishes as the targets approach the 100% proficiency requirement dictated under NCLB for 2014.**[3]

- In Rhode Island, as in most states, schools with fewer subgroups attain AYP more easily in Rhode Island than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As in other states, middle schools in Rhode Island have greater difficulty reaching AYP than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of a school making AYP under Rhode Island's system is whether it has enough English language learners to qualify as a separate subgroup. Every school with a subgroup of students with limited English proficiency (LEP)[4] failed to make AYP. Likewise, all but one school

[3] Rhode Island is one of six states studied (Massachusetts, Minnesota, Vermont, Wisconsin, and New Hampshire are the others) that uses an index that 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.

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

(King Richard) with enough qualifying SWDs failed to meet their AYP targets.[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Rhode Island's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state. Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008.)

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB

[5] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 New England Common Assessment Program (NECAP), the standardized state test. 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.

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

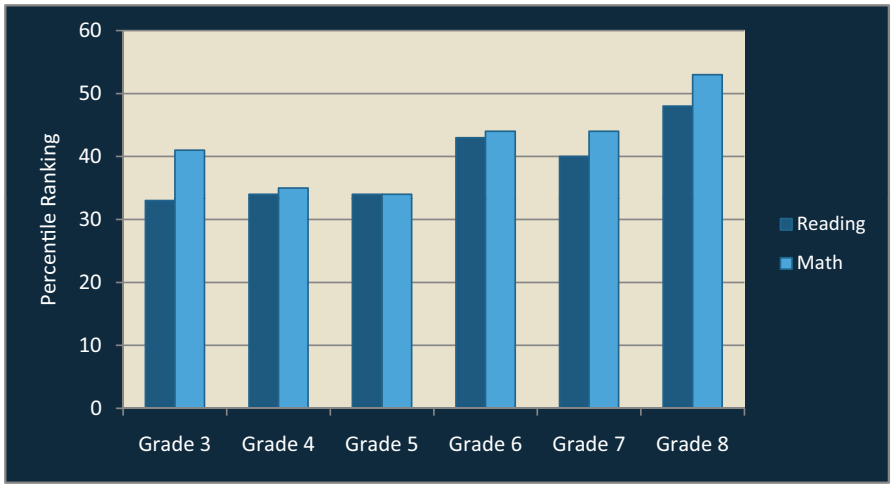

Figure 2. Rhode Island reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Rhode Island's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Rhode Island's cut scores are below the 55th percentile.

subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the New England Common Assessment Program (NECAP) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Rhode Island's definitions of proficiency generally ranked about average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Rhode Island test (NECAP), given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Rhode Island AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Rhode Island's 2008 AYP rules.

Table 1 shows the pertinent Rhode Island AYP rules that were applied to elementary and middle schools in this study. Rhode Island's minimum subgroup size is 45, which is slightly larger than the subgroup size for other states we examined.[8] This means that schools in Rhode Island may have fewer subgroups than similar schools in other states.

Rhode Island, like the majority of states examined, applies a 95% confidence interval to its measurements of student proficiency rates. This means even though the AMO might require a school to attain, for instance, 84.1% reading proficiency among its grade 3 students, and 84.1% reading proficiency among its grade 3 students in each subgroup, the real target can be lower, particularly with smaller groups.

Unlike most states, Rhode Island measures its student performance with a proficiency index, which gives partial credit for students achieving "partial proficiency." In the

[7] We gave all schools in our sample pseudonyms in this report.

[8] Keep in mind, however, that school size and *n* size are related (e.g., large *n* sizes make sense for larger schools).

Table 1. Rhode Island AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 45	
	SWDs: 45	
	Low-income students: 45	
	LEP students: 45	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (index)**	**2008 targets (index)**
READING/LANGUAGE ARTS		
Grade 3	76.1	84.1
Grade 4	76.1	84.1
Grade 5	76.1	84.1
Grade 6	68.0	78.6
Grade 7	68.0	78.6
Grade 8	68.0	78.6
MATH		
Grade 3	61.7	74.5
Grade 4	61.7	74.5
Grade 5	61.7	74.5
Grade 6	46.1	64.1
Grade 7	46.1	64.1
Grade 8	46.1	64.1

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

short term, the index makes it easier for Rhode Island schools to meet their targets, although the effect of the index diminishes as the targets approach the 100% proficiency requirement dictated under NCLB for 2014.[9]

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their

[9] In six of the states studied (Massachusetts, Minnesota, Vermont, New Hampshire, and Wisconsin, as well as Rhode Island), an index is used that 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.

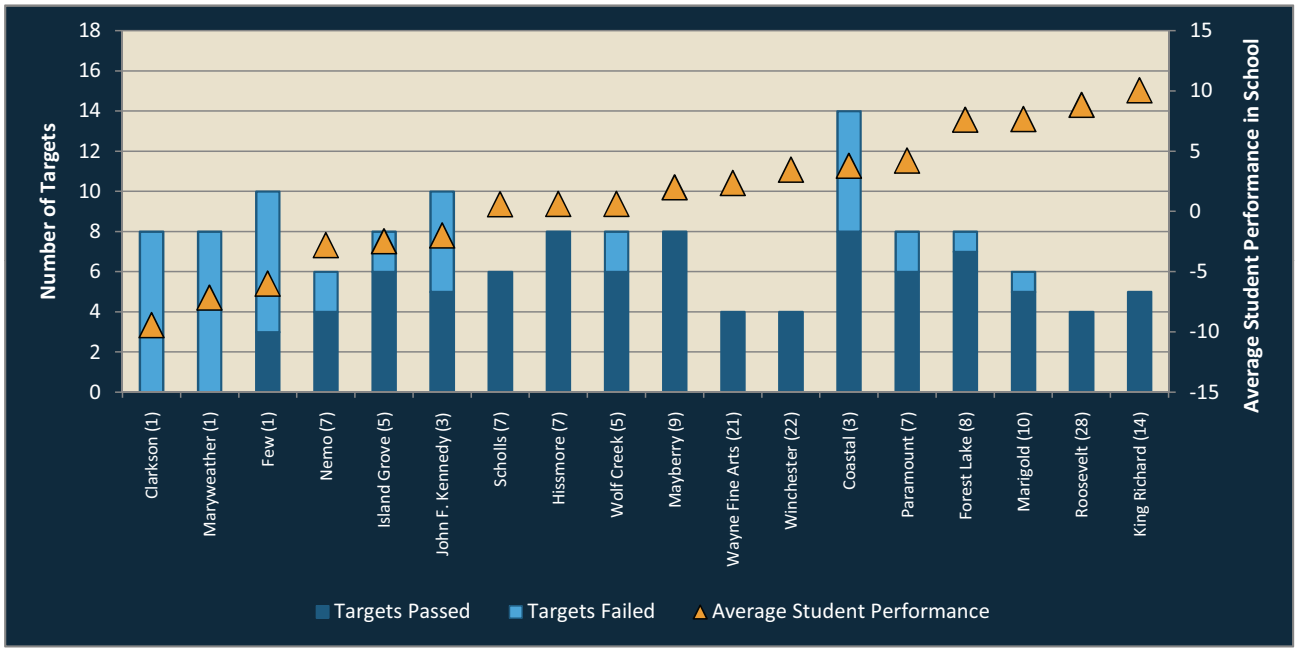

Figure 3. AYP performance of the elementary school sample under Rhode Island's 2008 AYP rules

Note: This figure shows how each elementary school within the sample fares under Rhode Island's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Marigold Elementary, for example, met five of its six targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students, and 95% of the students in each school's subgroup, to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Rhode Island's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Rhode Island's 2008 AYP rules. **Seven elementary schools (Scholls, Hissmore, Mayberry, Wayne Fine Arts, Winchester, Roosevelt, and King Richard) made AYP and 11 failed.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student and positive values indicating above-grade-level performance. For the most part, the schools that made AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet. For example, Scholls passed but has only six targets (two in reading and math for its overall school populations, two for its low income subgroups, and two for its white subgroups).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Rhode Island AYP rules. **Out of 18 middle schools in our sample, only 2 made AYP**—one low-performance school (Pogesto) and one high-performance school (Walter Jones), both of which

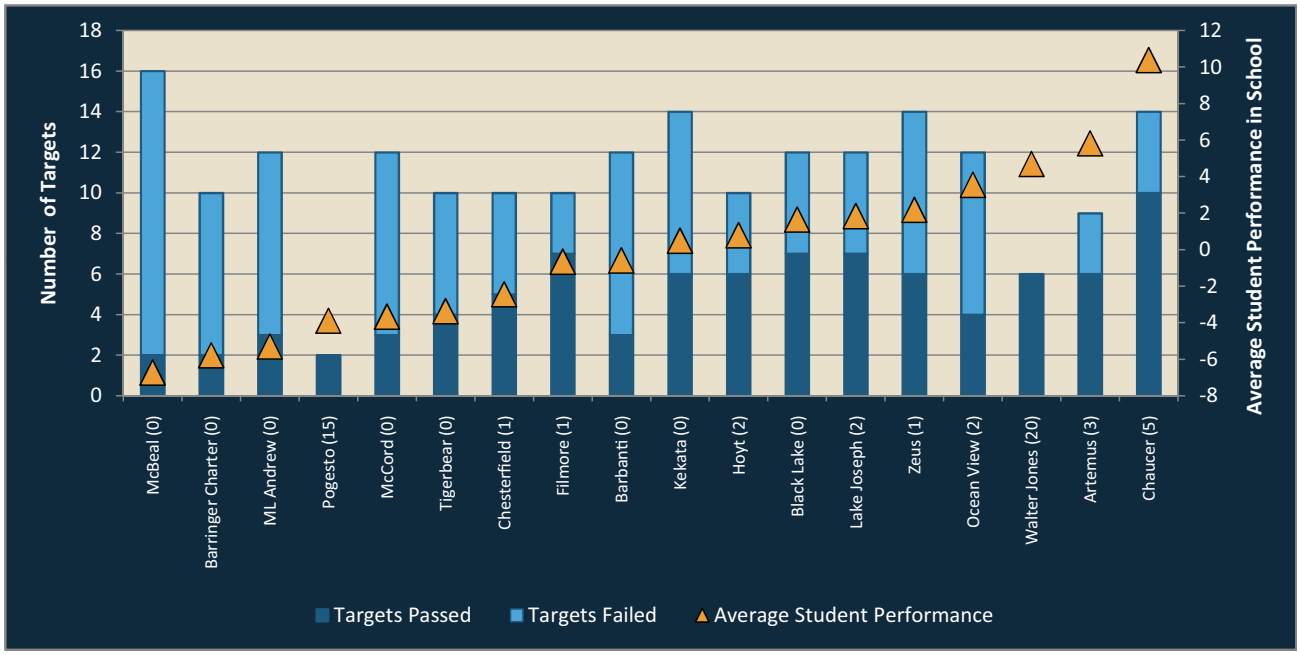

Figure 4. AYP performance of the middle school sample under Rhode Island's 2008 AYP rules

Note: This figure shows how each middle school within the sample would have fared under Rhode Island's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means that the school failed to make AYP. Artemus, for example, met 6 of its 9 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

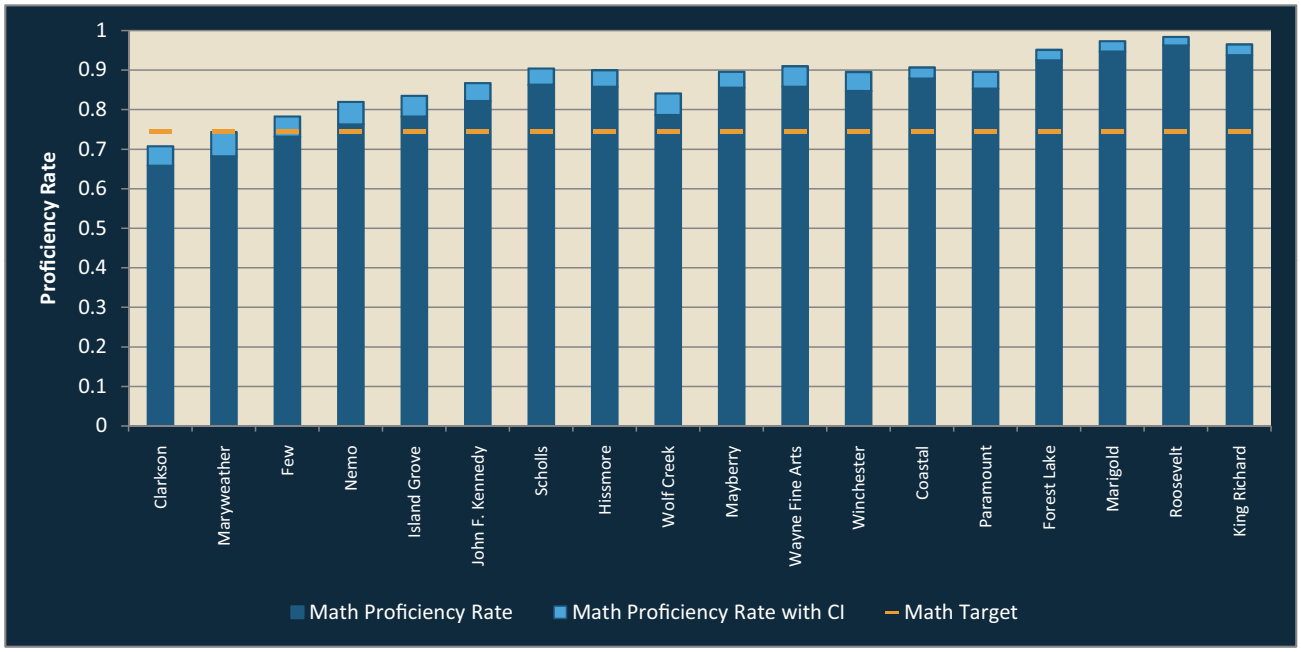

Figure 5. Impact of the confidence interval on elementary school mathematics proficiency rates under Rhode Island's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (Few) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

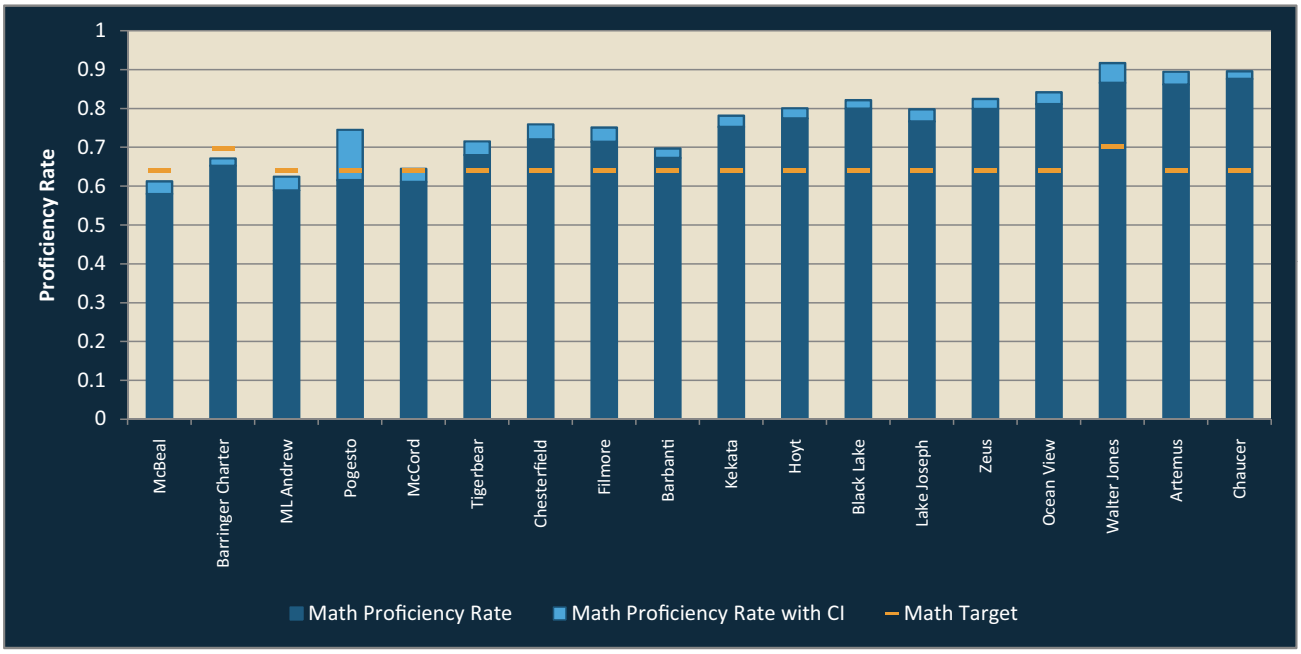

Figure 6. Impact of the confidence interval on middle school mathematics proficiency rates under Rhode Island's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools (Pogesto) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

have the lowest number of qualifying subgroups (and hence, targets to meet).

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates were increased by applying the confidence interval. The orange lines show the annual measurable objective needed to meet AYP. These figures show that one of the sample elementary schools (Few) and one middle school (Pogesto) are assisted by the confidence intervals, although we know from Figure 3 that Few Elementary still failed to make AYP because of subgroup performance. The effect of confidence intervals on schools' reading proficiency rates for elementary and middle schools is much the same (not shown).

In reading (not shown), four elementary schools (Nemo, Island Grove, Scholls, and Wolf Creek) and one middle school (Pogesto) are able to meet the overall target with the confidence interval, although we know from Figures 3 and 4 that most of these schools still failed to make AYP since they didn't meet targets for subgroups. **Overall, the application of the confidence interval appears to have modest impact on AYP decisions.**[10]

Where do schools fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still pass AYP when the school has fewer targets to meet because it has fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

[10] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Rhode Island AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	65.8%	62.0%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	68.2%	66.4%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Few	73.2%	69.7%	Y	N	N	N	N	N	Y	N					Y	N					10	3	30%	N	1
Nemo	76.3%	81.5%	Y	Y					N	N									Y	Y	6	4	67%	N	7
Island Grove	78.3%	79.9%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	4
JFK	82.2%	78.0%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	3
Scholls	86.3%	81.7%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Hissmore	85.7%	84.3%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Wolf Creek	78.7%	81.2%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	5
Alice Mayberry	85.5%	86.2%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	85.8%	92.4%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	84.7%	88.3%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	87.9%	83.4%	Y	Y	N	N	Y	N	Y	N	Y	N			Y	N			Y	Y	14	8	57%	N	3
Paramount	85.3%	84.6%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	92.4%	92.0%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	94.7%	91.2%	Y	Y					Y	N									Y	Y	6	5	83%	N	10
Roosevelt	96.2%	96.2%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	93.8%	93.5%	Y	Y	Y														Y	Y	5	5	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 Rhode Island rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Most schools met their targets for their overall school populations, except for two (Clarkson and Maryweather) that failed to meet the overall reading and math targets, and two others (Few and JFK) that failed to meet their overall reading targets.
- Three middle schools (McBeal, Barringer, and ML Andrew) failed to meet both their reading and math

Table 3. Middle school subgroup performance of sample schools under the 2008 Rhode Island AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	58.0%	65.0%	N	N	N	N	N	N	N	N	N	N			N	N	N	N	Y	Y	16	2	13%	N	0
Barringer Charter	65.3%	73.1%	N	N	N	N			N	N	N	N			Y	Y					10	2	20%	N	0
ML Andrew	58.9%	71.2%	N	N	N	N			N	N	N	N			Y	N			Y	Y	12	3	25%	N	0
Pogesto	61.6%	74.5%	Y	Y																	2	2	100%	Y	15
McCord Charter	61.1%	73.9%	Y	N	N	N			N	N	N	N			N	N			Y	Y	12	3	25%	N	0
Tigerbear	68.0%	68.8%	Y	N	N	N			Y	N	N	N							Y	Y	10	4	40%	N	0
Chesterfield	72.1%	72.9%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	1
Filmore	71.4%	78.9%	Y	Y	N	N			Y	Y					Y	N			Y	Y	10	7	70%	N	1
Barbanti	67.3%	74.1%	Y	N	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	75.3%	76.5%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	0
Hoyt	77.5%	79.3%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	2
Black Lake	80.0%	78.7%	Y	Y	N	N			Y	N	Y	N			Y	N			Y	Y	12	7	58%	N	0
Lake Joseph	76.7%	82.4%	Y	Y	N	N	N	N	Y	Y					Y	N			Y	Y	12	7	58%	N	2
Zeus	79.9%	80.5%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	1
Ocean View	81.1%	87.3%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	86.6%	88.9%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	86.2%	85.9%	Y	Y		N			Y	N					Y	N			Y	Y	9	6	67%	N	3
Chaucer	87.7%	91.0%	Y	Y	N	N	N	N	Y	Y			Y	Y	Y	Y			Y	Y	14	10	71%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

targets for their overall populations. Four others (McCord, Tigerbear, Chesterfield, and Barbanti) failed to meet their overall reading targets.

- One elementary school (Forest Lake) met its targets for every subgroup except SWDs.
- One elementary school (Marigold) met all its targets except for its low-income subgroup.
- No middle school met its targets for its SWD or students with LEP subgroups.

Tables 4 and 5 summarize the performance of the various subgroups for elementary and middle schools, respectively. Note, first, the performance of SWDs is proving quite challenging for schools under Rhode Island's system, particularly in middle schools, where these subgroups tend to have enough students to meet the state's minimum *n* of 45. In fact, all but one elementary (King Richard) and every middle school in the study with qualifying SWD subgroups failed to make AYP. Similarly, every single school with a population of students with LEP large enough to qualify as a separate subgroup failed to meet its reading targets for these students

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Rhode Island AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	5	3	4
Students with limited English proficiency	4	3	4
Low-income students	14	3	10
African-American students	4	0	2
Asian/Pacific Islander students	0	0	0
Hispanic students	7	2	7
American Indian/Alaska Native students	0	0	0
White students	15	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Rhode Island AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	15	15	16
Students with limited English proficiency	7	7	7
Low-income students	17	6	13
African-American students	10	5	10
Asian/Pacific Islander students	1	0	0
Hispanic students	13	6	11
American Indian/Alaska Native students	1	1	1
White students	16	0	0

and only one such school (Coastal) reached its math target for this subgroup.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Rhode Island's NCLB accountability system is, in many respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Rhode Island, too. Likewise, the elementary and middle schools that failed to make AYP in the greatest number of states also failed to make AYP in Rhode Island.

But Rhode Island is also home to a few anomalies. First,

Table 6. Comparisons between schools that did and didn't make AYP in Rhode Island, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	7	11	2	16
Average student body size	268	328	124	951
Average % low income	37	52	42	45
Average % nonwhite	32	47	27	46
Average performance†	3.99	-0.54	0.40	-0.11
Average % growth‡	112	117	109	97
Average number of targets to meet	6	9	4	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

consider Mayberry Elementary (see Figure 3). It failed to make AYP in 19 of the 28 states in our sample, yet made AYP in Rhode Island. Examining Table 2 we can see that Mayberry didn't meet the minimum numbers for the LEP or SWD subgroups, which create difficulty for so many other schools within the sample. With fewer accountable subgroups, Mayberry made meet AYP, even when other schools with higher average performance failed. Second, look at Pogesto Middle School (Figure 4). Even with its relatively low average performance, it made AYP in Rhode Island, while failing to do so in 13 of 28 states. Like Mayberry, its AYP success in Rhode Island is likely attributable to the relatively small number of targets (two) it has to meet, as shown in Table 3.

This is consistent with the patterns shown in Table 6, which compares schools making and not making AYP on a number of academic and demographic dimensions. Within the sample, schools that made AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations (especially at the middle school level), fewer subgroups (and thus fewer targets to meet), and lower percentages of non-white students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Rhode Island's AYP rules and AMOs for 2008. We found that 7 elementary schools and 2 middle schools—9 in all from a total of 36—would have made AYP in Rhode Island. Looking across the 28 state accountability systems examined in the study, this puts Rhode Island in the upper middle of the distribution in terms of the number of schools making AYP (as shown in Figure 1). Rhode Island's proficiency standards (or cut scores) are above average compared to other states; similarly, the state's annual targets for proficiency are fairly high, particularly at the elementary school level. However, Rhode Island uses a proficiency index, which gives partial credit for students achieving "partial proficiency." In the short term, such an index makes it easier for Rhode Island schools to meet their targets, although the effect of the index diminishes as the targets approach the 100% proficiency requirement dictated under NCLB for 2014.

Because the overriding goal of the federal NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Rhode Island's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Most of the sample schools met the Rhode Island math and reading targets for their student populations as a whole. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils aren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Similarly, does it make sense that subgroup participation differs so much between elementary and middle schools, as it does in Rhode Island? Is it "fair" that, in Rhode Island and in a handful of other states, students are awarded "partial" credit even though they did not achieve proficiency? And equally important, doesn't the failure of English language learners and SWDs[11] to meet Rhode Island's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for students with LEP and SWDs, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

[11] See footnote 5.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

South Carolina

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines South Carolina's NCLB accountability system—particularly how its various rules, criteria and practices result in schools either making AYP—or not making AYP. It also gauges how tough South Carolina's system is compared with other states. For this study we selected 36 schools from around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under South Carolina's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against South Carolina's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **15 of 18 elementary schools** and **all 18 middle schools** in our sample **failed to make AYP** in 2008 under South Carolina's accountability system.
- This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students. But it's also partly explained by South Carolina's ambitious proficiency standards (or cut scores), which are among the most rigorous in our state sample.
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in South Carolina is exceeded in 15 other sample states** (South Carolina ties four other states with only three elementary schools making AYP). In addition, South Carolina is one of five states with no passing middle schools in the sample (see Figure 1).
- Unlike many other states in the study, most schools in South Carolina fail to meet math and reading targets for their overall populations. Again, this is likely due to the state's difficult proficiency standards.
- In South Carolina, as in most states, schools with fewer subgroups attain AYP more easily than schools

South Carolina's accountability system has several unique characteristics that land it in the middle of the state distribution in terms of the number of schools making AYP. First, South Carolina's cut scores are relatively difficult to achieve compared to other states in the study. Most all of them are above the 50th percentile and some are even around the 70th percentile. However, South Carolina adds the equivalent of one standard error to individual test scores, which essentially lowers the difficulty of the proficiency cut score (other states apply confidence intervals which have a similar effect). South Carolina also utilizes different subgroup sizes for different groups. The minimum subgroup size is 40 for racial, ethnic, and low-income groups (which is fairly standard), and 50 for students with disabilities (SWDs) and for students with limited English proficiency (LEP) (which is higher than in most other states). The latter means that fewer students are held separately accountable for performance in South Carolina that would similar schools in other states.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the South Carolina Palmetto Achievement Challenge Tests (PACT).

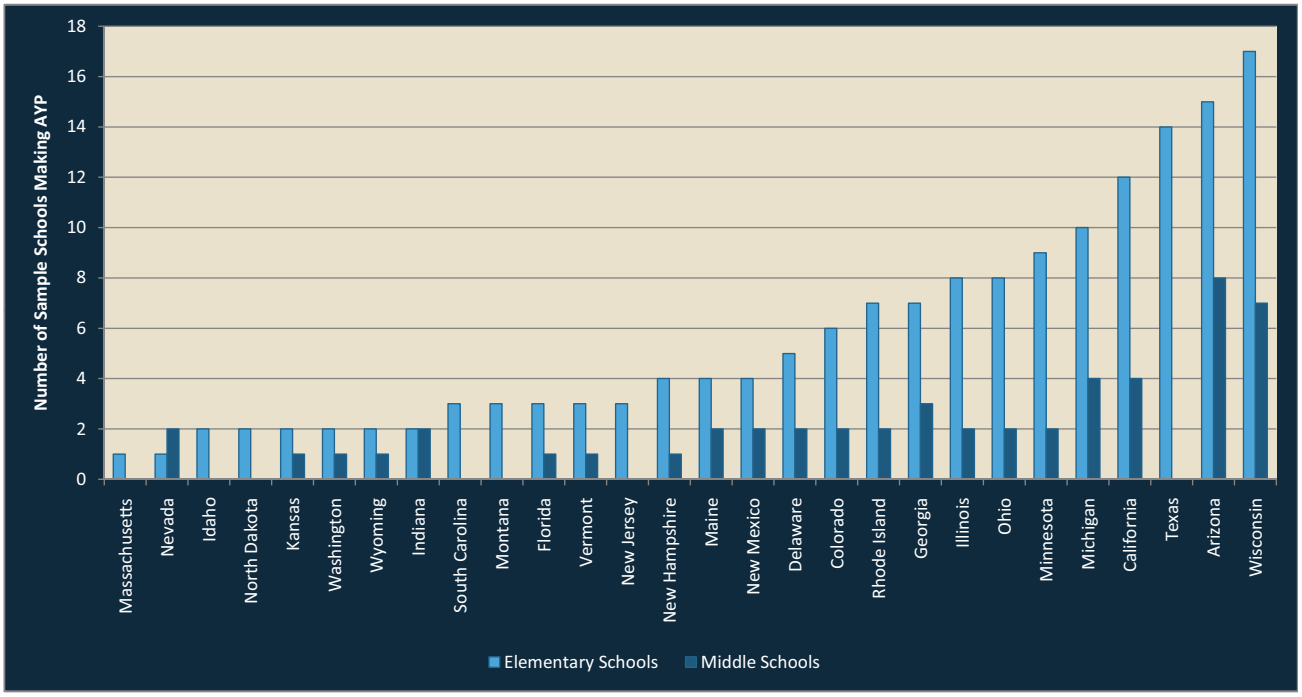

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As is the case in other states, middle schools in South Carolina have greater difficulty reaching AYP than do elementary schools, primarily because their student populations are larger and therefore, have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of a school making AYP under South Carolina's system is whether it has enough qualifying students with disabilities (SWDs) and students with limited English proficiency (LEP). **South Carolina has a large "minimum *n* size" (50) for these two groups, meaning schools in the state may have fewer accountable subgroups than would similar schools in other states.[2] Still, when enough students exist to comprise these subgroups, every single school with a SWD or LEP subgroup failed to make AYP**. Likewise, all but one school with enough qualifying low-income students failed to meet its AYP targets.[3]

Introduction

The Proficiency Illusion (Cronin et al.2007a) linked student performance on South Carolina's tests and those of 25 other state tests to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP),

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

[3] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, mostly because of the likely differences between how LEP students and SWDs are treated in the Measures of Academic Progress (MAP), the assessment we used in this study, and in the South Carolina Palmetto Achievement Challenge Tests (PACT), the standardized state test. 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.

a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether schools make AYP. Specifically, we estimate how a single set of schools, drawn from around the country, would fare under the differing rules for determining Adequate Yearly Progress (AYP) in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–14. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low-income,[4] African-American, among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school making AYP in Wisconsin or Ohio, for example, might not make it under Nevada's or Idaho's rules (U.S, Department of Education, 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners,[5] among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

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

[5] Note that we use "students with limited English proficiency (LEP)" or "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup.

[6] We gave all schools in our sample pseudonyms in this report.

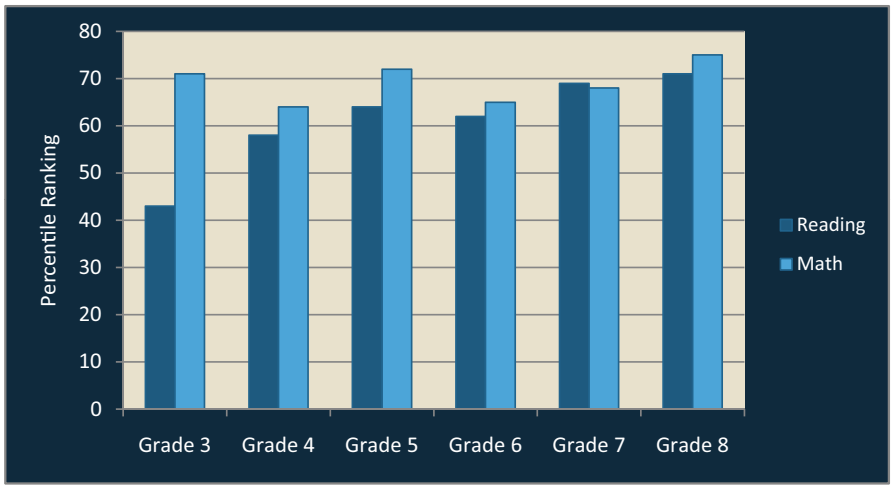

Figure 2. South Carolina reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of South Carolina's cut scores (or proficiency passing scores) for its reading and mathematics tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of South Carolina's cut scores are at or below the 75th percentile.

Proficiency cut score estimates for the South Carolina Palmetto Achievement Challenge Tests (PACT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that South Carolina's definitions of proficiency generally ranked well above the average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the South Carolina test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under South Carolina AYP rules for 2008. (In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against South Carolina's 2008 AYP rules.)

Table 1 shows the pertinent South Carolina AYP rules that were applied to elementary and middle schools in this study. South Carolina's minimum subgroup size is 40 for race/ethnicity and low-income groups, and 50 for students with disabilities and for students with limited English proficiency. While 40 is roughly comparable to the number used by most other states in the current study, 50 is a bit larger.[7] This means that schools in South Carolina may have fewer accountable subgroups than would similar schools in other states. Furthermore, while the majority of states examined in the study apply confidence intervals to their measurements of student proficiency rates, South Carolina adds the equivalent of one standard error to individual test scores, essentially lowering the difficulty of the proficiency cut score.[8]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are

[7] Keep in mind, however, that school size and *n* size are related (e.g., a small *n* size make sense for small schools).

[8] By adding a standard error to individual student scores, South Carolina essentially adds a few points to the score and effectively lowers the proficiency standard (or cut score). This is done to correct for potential measurement error of the state testing instrument, which is the same argument used by other states that use confidence intervals when reporting school or group proficiency rates. For the stated purpose (i.e., correcting for measurement error), South Carolina's approach is probably more appropriate. However, the current study did not systematically examine whether standard errors or confidence intervals provided greater assistance to schools.

Table 1. South Carolina AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 50	
	Low-income students: 40	
	LEP students: 50	
CI	**Applied to proficiency rate calculations?**	
	Not used, but one standard error added to individual test scores	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	17.6	58.8
Grade 4	17.6	58.8
Grade 5	17.6	58.8
Grade 6	17.6	58.8
Grade 7	17.6	58.8
Grade 8	17.6	58.8
MATH		
Grade 3	15.5	57.8
Grade 4	15.5	57.8
Grade 5	15.5	57.8
Grade 6	15.5	57.8
Grade 7	15.5	57.8
Grade 8	15.5	57.8

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under South Carolina's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under South Carolina's 2008 AYP rules. **Only 3 elementary schools made AYP while 15 failed to make it**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. All schools making AYP are in the right half of the figure, meaning

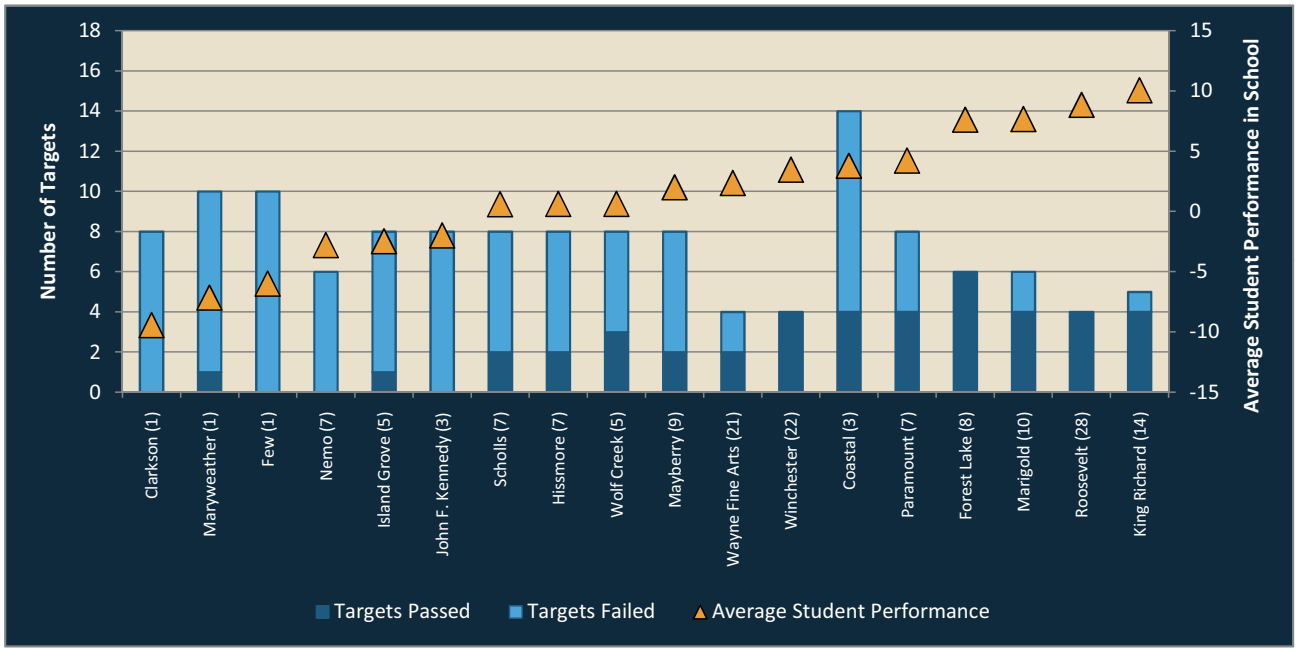

Figure 3. AYP performance of the elementary school sample under the South Carolina 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under South Carolina's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means the school failed. Marigold Elementary, for example, met 4 of its 6 targets, but because it did not meet them all, it did not make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles) which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school would have made AYP in the study.

that the highest performing students were found at these schools.

Yet almost without regard to average student performance, the only schools to make AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has separate targets). For example, Winchester made it, but it had only four targets –two in reading and math for its overall population, and two more for the only subgroup to exceed the minimum size (the white subgroup).

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 South Carolina AYP rules. **None of the sample middle schools made AYP**.

Where do schools fail?

Figures 3 and 4 indicate the number of subgroup targets evaluated in each sample school, and each school's final AYP outcome. However, these figures do not indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and White. Tables 2 and 3 also show whether a school met AYP under the 2008 South Carolina rules, and the total number of states within the study in which that school met AYP.

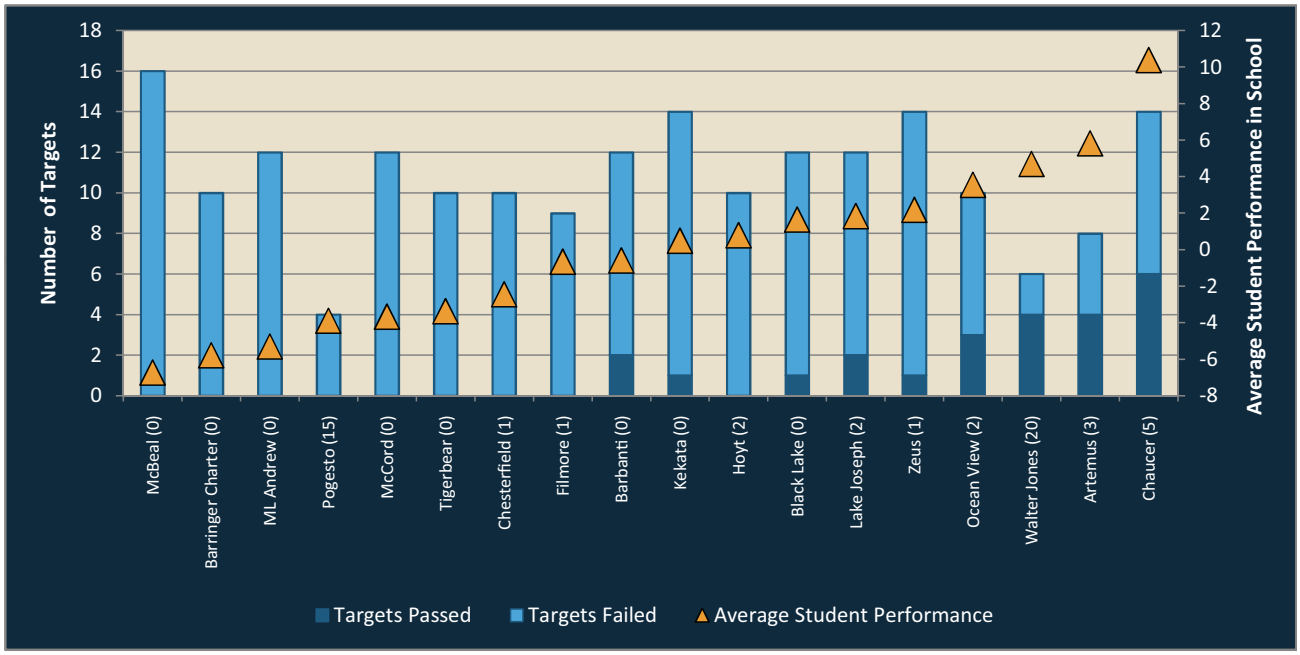

Figure 4. AYP performance of the middle school sample under the South Carolina 2008 AYP rules

Note: This figure indicates how each of the middle schools within the sample fared under South Carolina's AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup does not make AYP, so any light blue means the school failed. Walter Jones, for example, met 4 of its 6 targets, but because it did not meet them all, it did not make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school would make AYP in the study.

The school-by-school findings in Tables 2 and 3 show that:

- In South Carolina, most schools failed to meet math and reading targets for their overall populations, unlike many other states in the study.
- Seven elementary schools meet both the reading and math targets for their overall populations.
- Three middle schools met both the reading and math targets for their overall populations.
- Overall, in most of the cases where there were enough students to comprise a subgroup, these groups didn't meet their targets.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, one can see that nearly all subgroups at both the elementary and middle school levels struggle with South Carolina's reading and math requirements, perhaps because South Carolina's proficiency cut scores are very difficult compared to the other states in the sample. Every school with sufficient numbers of SWDs, LEP students, low income students, African American, and Hispanic subgroups failed to make AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that South Carolina's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt and Winchester made AYP in the greatest number of states—28 and 22 respectively. And these schools made AYP in South Carolina, too. Likewise, the elementary and middle schools that fail to make AYP in the greatest number of states also failed to make AYP in South Carolina.

But South Carolina is home to at least one anomaly. First, consider Forest Lake Elementary (see Figure 3). It failed to make AYP in 20 of the 28 states in our sample,

Table 2. Elementary school subgroup performance of sample schools under the 2008 South Carolina AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	27.0%	23.1%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	30.6%	38.4%	N	N			N	N	N	N					N	N			N	Y	10	1	10%	N	1
Few	36.1%	37.8%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Nemo	35.3%	53.5%	N	N					N	N									N	N	6	0	0%	N	7
Island Grove	39.9%	56.8%	N	N					N	N					N	N			N	Y	8	1	13%	N	4
JFK	43.1%	48.5%	N	N					N	N	N	N							N	N	8	0	0%	N	3
Scholls	54.1%	56.1%	N	N					N	N	N	N							Y	Y	8	2	25%	N	7
Hissmore	53.2%	58.4%	N	N					N	N	N	N							Y	Y	8	2	25%	N	7
Wolf Creek	55.0%	58.9%	N	Y					N	N					N	N			Y	Y	8	3	38%	N	5
Alice Mayberry	54.1%	57.4%	N	N					N	N	N	N							Y	Y	8	2	25%	N	9
Wayne Fine Arts	48.3%	67.2%	N	Y															N	Y	4	2	50%	N	21
Winchester	58.0%	68.7%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	64.8%	61.5%	Y	Y	N	N	N	N	N	N	N	N			N	N			Y	Y	14	4	29%	N	3
Paramount	68.8%	67.9%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	7
Forest Lake	76.7%	76.6%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	8
Marigold	75.9%	75.1%	Y	Y					N	N									Y	Y	6	4	67%	N	10
Roosevelt	74.7%	83.4%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	76.0%	81.3%	Y	Y					N										Y	Y	5	4	80%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

yet made AYP in South Carolina. Examining Table 2, one can see that Forest Lake didn't meet the minimum numbers for the LEP or SWD subgroups, perhaps because South Carolina's minimum "*n*" for these groups is higher than in most of the other states examined. With fewer accountable subgroups, Forest Lake made AYP, even when other schools with higher average performance failed.

That fewer accountable subgroups is a good predictor of making AYP is consistent with the patterns shown in Table 6, which compares elementary schools (there were no passing middle schools) that made and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that made AYP do indeed show higher average student performance, but they also differ in the following ways: they have smaller student populations, fewer subgroups (and thus fewer targets to meet), and lower percentages of low income students.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the

Table 3. Middle school subgroup performance of sample schools under the 2008 South Carolina AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	28.9%	33.7%	N	N	N	N	N	N	N	N	N	N			N	N	N	N	N	N	16	0	0%	N	0
Barringer Charter	28.5%	37.2%	N	N	N	N			N	N	N	N			N	N					10	0	0%	N	0
ML Andrew	26.4%	34.3%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Pogesto	29.6%	20.4%	N	N															N	N	4	0	0%	N	15
McCord Charter	32.3%	40.1%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Tigerbear	38.9%	34.9%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	0
Chesterfield	40.0%	33.5%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	1
Filmore	40.4%	42.9%	N	N		N			N	N					N	N			N	N	9	0	0%	N	1
Barbanti	42.4%	43.3%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Kekata	49.7%	43.9%	N	N	N	N	N	N	N	N	N	N			N	N			Y	N	14	1	7%	N	0
Hoyt	45.1%	46.6%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	2
Black Lake	52.5%	43.2%	N	N	N	N			N	N	N	N			N	N			Y	N	12	1	8%	N	0
Lake Joseph	49.1%	49.4%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	2
Zeus	53.7%	50.2%	N	N	N	N	N	N	N	N	N	N			N	N			Y	N	14	1	7%	N	1
Ocean View	52.8%	58.9%	N	Y			N	N	N	N					N	N			Y	Y	10	3	30%	N	2
Walter Jones	63.4%	70.3%	Y	Y					N	N									Y	Y	6	4	67%	N	20
Artemus	66.2%	63.2%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	3
Chaucer	68.4%	70.7%	Y	Y	N	N	N	N	N	N			Y	Y	N	N			Y	Y	14	6	43%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

country to see how these schools would fare under South Carolina's AYP rules and annual measurable objectives for 2008. We found that, only three elementary schools and no middle schools–three in all from a total of 36–would have made AYP in South Carolina. Looking across the 28 state accountability systems examined in the study, this puts South Carolina at the low end of the distribution in terms of the number of elementary schools making AYP (see Figure 1). Part of the reason that so few schools make AYP in South Carolina may be due to its ambitious proficiency standards, which are among the most rigorous in our state sample. Indeed, unlike many other states in the study, most schools in South Carolina fail to meet math and reading targets for their overall populations.[9]

The overriding goal of the No Child Left Behind act (NCLB) is to eliminate educational disparities within and across states, it's important to consider whether

[9] It does not appear that South Carolina's high proficiency standards have had much impact on the state's performance on the latest (2007) National Assessment of Educational Progress (NAEP). South Carolinian children performed lower than the national average in grade 4 and 8 reading, as well as in grade 4 math.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 South Carolina AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	2	2	2
Students with limited English proficiency	4	4	4
Low-income students	15	14	13
African-American students	5	5	5
Asian/Pacific Islander students	0	0	0
Hispanic students	7	7	7
American Indian/Alaska Native students	0	0	0
White students	16	5	2

Table 5. Summary of subgroup performance of sample middle schools under the 2008 South Carolina AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	14	13	14
Students with limited English proficiency	7	7	7
Low-income students	17	17	17
African-American students	10	10	10
Asian/Pacific Islander students	1	0	0
Hispanic students	13	13	13
American Indian/Alaska Native students	1	1	1
White students	17	8	11

states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, South Carolina's No Child Left Behind accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils aren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a good thing.

Yet NCLB's design flaws are also readily apparent. Does it make sense, in the case of South Carolina, that schools

Table 6. Comparisons between schools that did and didn't make AYP in South Carolina, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	3	15	0	18
Average student body size	243	317	n/a	859
Average % low income	20	52	n/a	45
Average % nonwhite	21	45	n/a	44
Average performance†	6.65	0.14	n/a	-0.05
Average % growth‡	131	112	n/a	98
Average number of targets to meet	5	8	n/a	11

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

n/a = not applicable

are penalized for the state's high proficiency standards? Does it make sense that fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[10] doesn't the failure of these students (especially at the middle school level where more satisfy eligibility requirements) to meet South Carolina's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP and SWD students, as for other students, but when so few schools are able to meet the goal perhaps that indicates the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

[10] See footnote 3.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Texas

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Texas's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP—or not making AYP. It also gauges how tough Texas's system is compared with those of other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Texas's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Texas's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **4 of 18 elementary schools** in our sample **failed to make AYP** in 2008 under Texas's accountability system.
- Looking **across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in Texas was exceeded in just 2 other sample states (Arizona and Wisconsin)**. (Note that middle schools were not examined in Texas, unlike other states, since eighth grade cut scores were not available.)
- Part of the reason that so many schools make AYP in Texas is that its **proficiency standards are relatively easy, compared to other states. Schools also have fewer accountable subgroups in Texas, likely because the state has a relatively large minimum "*n* size" for holding subgroups accountable.**
- Nearly all the schools in our sample that failed to make AYP in Texas are meeting expected targets for their overall populations[2] but failing because of the performance of individual subgroups, particularly students with disabilities (SWDs) and students with limited English proficiency (LEP).[3]

Just four of 18 elementary schools in our sample fail to make AYP in 2008 under **Texas's** accountability system. Looking across the 28 state accountability systems examined in the study, we find Texas to be among the least restrictive in terms of how many sample schools make AYP. This is likely due to a number of factors. First, Texas's proficiency standards (or cut scores) are relatively easy. Almost all of Texas's cut scores are below the 35th percentile. Second, Texas has a relatively large minimum *n* size for subgroup reporting, meaning that schools in Texas will have fewer accountable subgroups than would similar schools in other states. Unlike most other states, though, Texas does not report a confidence interval around its proficiency rates, but we generally found that they had limited impact on schools' AYP status in the study anyway.

[1] A cut score is the minimum score a student must receive on the Texas Assessment of Knowledge and Skills in order to be considered proficient under Texas's accountability system.

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup.

[3] SWDs are defined as those students following individualized education plans. Also, note that we use "LEP students" and "English language learners" interchangeably to refer to students in the same subgroup.

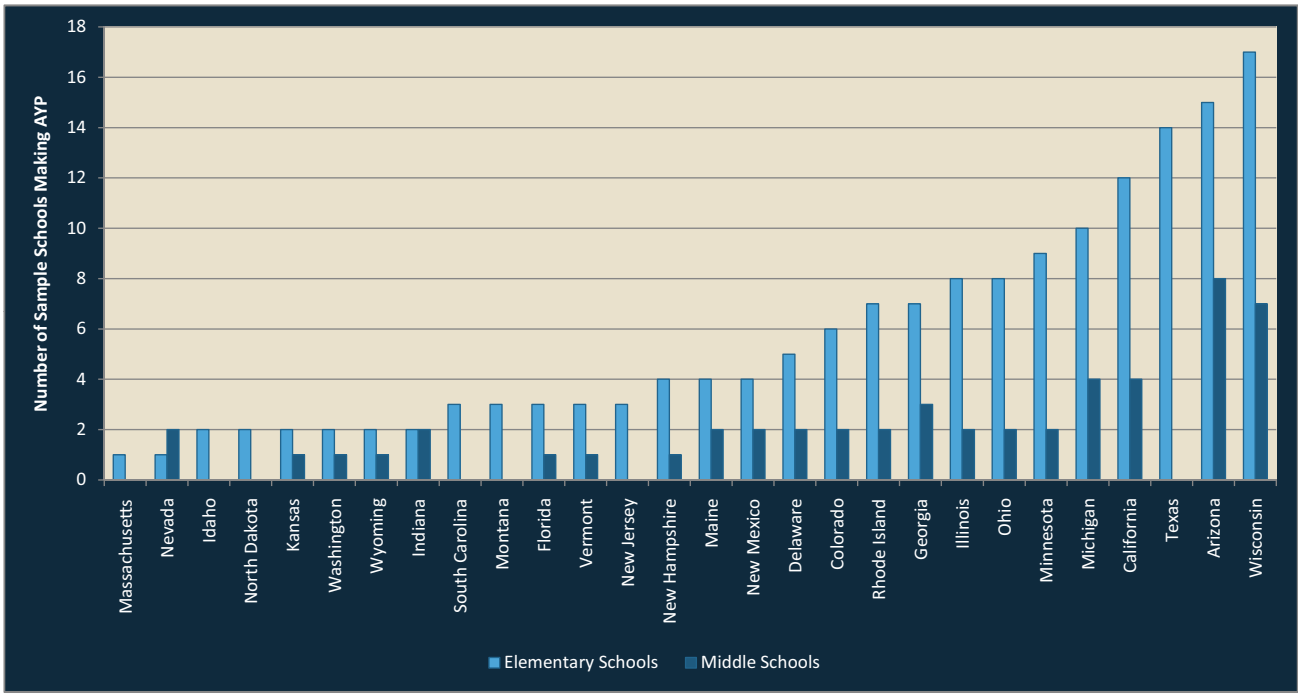

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- Ten sample elementary schools that failed to make AYP in most other states made AYP in Texas. Again, this is likely due to the state's easy proficiency standards and large minimum subgroup size.
- In Texas, as is the case in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.
- A strong predictor of a school making AYP under Texas's system is whether it has enough SWDs or LEP students to qualify as a separate subgroup. Every single school with these subgroups failed to make AYP.[4]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Texas's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability

[4] It should be noted that our subgroup findings for Limited English Proficient (LEP) and students with disabilities may be slightly more negative than would be seen under real world conditions. This is mostly due to the differences in testing practices between how LEP students and students with disabilities are treated in the Texas Assessment of Knowledge and Skills (TAKS) state assessment and in the NWEA's Measures of Academic Progress (MAP), the assessment used in this study. Specifically, the U.S. Department of Education has issued NCLB guidelines permitting schools to exclude small percentages of LEP or disabled students from taking state tests, or providing them alternate assessments. In the current study, however, no valid MAP scores were omitted from consideration.

plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008.)

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the Texas Assessment of Knowledge and Skills (TAKS) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Texas's definitions of proficiency were below the average, or less difficult, compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Texas test, given their performance on MAP. Student test data and subgroup designations are then used to determine how these 18 elementary schools would have fared under Texas AYP rules for 2008. In

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

[6] We gave all schools in our sample pseudonyms in this report.

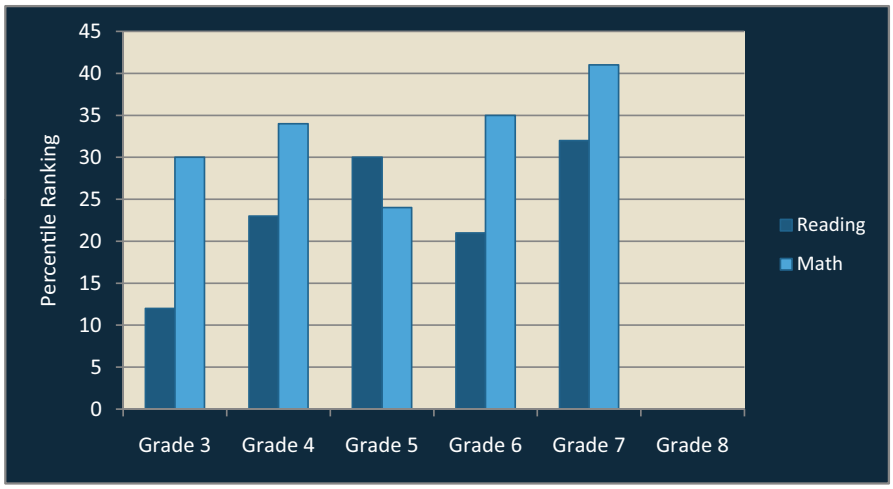

Figure 2. Texas reading and math cut score estimates, expressed as percentile ranks (2006))

Note: This figure illustrates the difficulty of Texas's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Texas's cut scores are below the 45th percentile. Cut scores for eighth grade were not available.

other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Texas's 2008 AYP rules. Note that in Texas, unlike most of the other state reports, the 18 sample middle schools were not examined, since the eighth grade cut scores were not available for Texas. Consequently, for Texas, only the performance of the sample elementary schools was examined.

Table 1 shows the pertinent Texas AYP rules that were applied to elementary schools in this study. Texas's minimum subgroup size is 10% of the population, if that is at least 50 but not more than 200.[7] This is a larger subgroup size than in many of the other states examined, meaning that schools in Texas will have fewer accountable subgroups than would similar schools in other states. Unlike most of the states in the study, Texas does not report a confidence interval around its proficiency rates. This means that schools in Texas will have greater difficulty achieving their targets than would schools that do use confidence intervals.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students—and 95% of the students in each school's subgroup—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We

[7] In Texas, the minimum subgroup size is 10% of the total school population. Generally, this means that the subgroup size grows with the school size. However, there's also a clause that specifies the minimum subgroup size can't be less than 50 or more than 200. For example, a school with a total population of 1000 would have a minimum subgroup size of 100 (i.e., 10%), but a school with only 400 students would have a minimum subgroup size of 50, since 10% of 400 (i.e., 40) is below the minimum. Similarly, a school with 3,000 students would have a minimum subgroup size of 200, since 10% of 3,000 (i.e., 300) is greater than the maximum value.

Table 1. Texas AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 10% of school population if at least 50 but not more than 200	
	SWDs: 10% of school population if at least 50 but not more than 200	
	Low-income students: 10% of school population if at least 50 but not more than 200	
	LEP students: 10% of school population if at least 50 but not more than 200	
CI	**Applied to proficiency rate calculations?**	
	Not used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	46.8	60.0
Grade 4	46.8	60.0
Grade 5	46.8	60.0
Grade 6	46.8	60.0
Grade 7	46.8	60.0
Grade 8	46.8	60.0
MATH		
Grade 3	33.4	50.0
Grade 4	33.4	50.0
Grade 5	33.4	50.0
Grade 6	33.4	50.0
Grade 7	33.4	50.0
Grade 8	33.4	50.0

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Texas's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Texas's 2008 AYP rules. **Fourteen elementary schools made AYP while only 4 failed to make it.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. Three of the schools not making AYP (Clarkson, Maryweather and Few) are in the left half of the figure, meaning that the lowest performing students were found at these schools.

Yet almost without regard to average student performance, the schools that failed to make AYP were those with relatively more qualifying subgroups—and thus the most targets to meet (because each subgroup has separate targets). For example, Coastal has relatively high performing students when compared to the other schools in the sample. However, it has the highest number of targets (12) and did not make AYP; whereas, Nemo is a school with lower performing students and made AYP, likely due to the low number of targets (6).

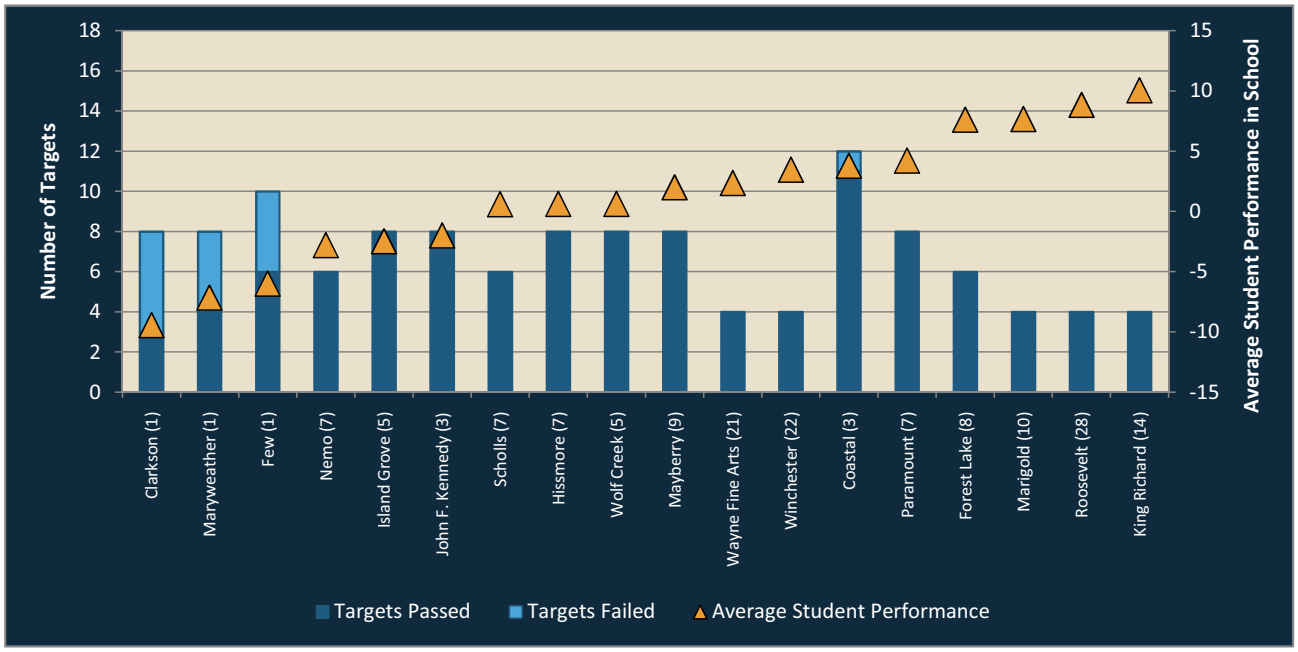

Figure 3. AYP performance of the elementary school sample under Texas's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Texas's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Coastal Elementary, for example, meets 11 of its 12 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

Where do schools fail?

Figure 3 illustrates that schools with low or middling performance can still make AYP when the school has fewer targets to meet because it has fewer subgroups. This figure does not indicate which subgroups failed or passed in which school. Table 2 lists information on individual subgroup performance.

Table 2 shows which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Table 2 also shows whether a school met AYP under the Texas rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 show that:

- Only 2 schools have enough SWDs to comprise a separate subgroup. Only three schools have enough LEP students to comprise a separate subgroup. None of these schools made AYP.
- One elementary school (Clarkson) failed to meet the reading targets for its overall school population. No elementary schools failed to meet their overall math targets.
- One failing elementary school (Coastal) met its targets for every subgroup except for SWDs.
- All low income subgroups met their math targets.

Table 3 summarizes the performance of the various subgroups. First, the performance of LEP students is proving challenging for schools under Texas's system; all three schools with large enough LEP populations to qualify as

Table 2. Elementary school subgroup performance of sample schools under the 2008 Texas AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	56.2%	56.1%	Y	N			N	N	Y	N					Y	N					8	3	38%	N	1
Maryweather	59.8%	62.1%	Y	Y			N	N	Y	N					Y	N					8	4	50%	N	1
Few	69.1%	66.3%	Y	Y	N	N	N	N	Y	Y					Y	Y					10	6	60%	N	1
Nemo	68.8%	80.5%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	72.7%	77.0%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	4
JFK	75.5%	73.9%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	3
Scholls	82.8%	78.8%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Hissmore	82.5%	82.8%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Wolf Creek	72.9%	79.0%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	5
Alice Mayberry	82.4%	83.7%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	83.3%	91.4%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	82.1%	86.3%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	84.9%	79.4%	Y	Y	Y	N			Y	Y	Y	Y			Y	Y			Y	Y	12	11	92%	N	3
Paramount	82.9%	82.5%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	7
Forest Lake	90.9%	90.3%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	8
Marigold	92.8%	89.2%	Y	Y															Y	Y	4	4	100%	Y	10
Roosevelt	95.6%	96.3%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	90.5%	91.5%	Y	Y															Y	Y	4	4	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

separate subgroups fail to meet their reading and math targets for these students. SWDs are also struggling to meet the state's targets. Neither of the two schools with qualifying SWD subgroups made AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figure 3 indicates that Texas's NCLB accountability system is, in some respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Texas, too. But Texas is also home to quite a few anomalies. First, consider JFK Elementary School (Figure 3). Even with its relatively low average performance it made AYP in Texas, but failed to do so in 25 of 28 states. Its AYP success in Texas is most likely attributable to its relatively small number of targets under Texas's minimum subgroup size rule (see Table 2), along with Texas's relatively easy proficiency cut scores, compared to other states.

Table 3. Summary of subgroup performance of sample elementary schools under the 2008 Texas AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	2	1	2
Students with limited English proficiency	3	3	3
Low-income students	13	0	2
African-American students	4	0	0
Asian/Pacific Islander students	0	0	0
Hispanic students	7	0	2
American Indian/Alaska Native students	0	0	0
White students	15	0	0

Table 4. Comparisons between schools that did and didn't make AYP in Texas, 2008

	Elementary Schools	
	Made AYP	**Failed to make AYP**
Number of schools in sample	14	4
Average student body size	281	387
Average % low income	37	79
Average % nonwhite	31	76
Average performance†	2.92	-4.69
Average % growth‡	117	109
Average number of targets to meet	6	10

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

This is consistent with the patterns shown in Table 4, which compares schools that made and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower percentages of low income students.

Concluding Observations

This study examined the test performance data of students

from 18 elementary and 18 middle schools across the country to see how these schools would fare under Texas's AYP rules and annual measurable objectives for 2008. Among this sample, 14 elementary schools in Texas—14 from an elementary school sample of 18—would have made AYP in Texas (this study did not include examination of Texas middle schools). Looking across the 28 state accountability systems examined in the study, this puts Texas at the high end of the distribution in terms of the number of schools making AYP (see Figure 1). **The fairly large number of schools making AYP in Texas may be due to the fact that Texas's proficiency standards are relatively easy, compared to other states and because the state has a relatively large minimum *n* size for subgroup reporting, meaning fewer groups are held accountable than might be the case in other states.**[8] In fact, only two schools have enough SWDs to comprise a separate subgroup and only three schools have enough students with LEP to comprise a separate subgroup.

Because the overriding goal of the federal NCLB is to eliminate educational disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Texas's No Child Left Behind accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students such as low-income students. All but one of the sample schools met the Texas reading and math targets for their student populations as a whole. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Is it "fair," in Texas's case, that so few SWDs and students with LEP are counted separately, meaning schools have to meet fewer targets? And in the rare cases when they do count separately, that they consistently fail to meet their annual targets? These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in

[8] Keep in mind, however, that school size and *n* size are related (larger *n* sizes may make sense for larger schools).

a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Vermont

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Vermont's NCLB accountability system—particularly how its various rules, criteria and practices result in schools either making AYP—or not making AYP. It also gauges how tough Vermont's system is compared with other states. For this study, we selected 36 schools from around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether or not each would make AYP under Vermont's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Vermont's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **15 of 18 elementary schools** and **17 of 18 middle schools** in our sample fail to make adequate yearly progress in 2008 under Vermont's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students. But **it's also partly explained by Vermont's annual proficiency targets, which are fairly rigorous (roughly 87 percent of Vermont's grade 3-8 students are expected to be proficient in reading in 2008)**.

- Looking across the 28 state accountability systems examined in the study, we find **Vermont at about the middle of the distribution in terms of the number of elementary sample schools making AYP**. Specifically, it exceeds fifteen states and ties with four others (South Carolina, Montana, Florida and New Jersey) (See Figure 1).

- Some of the schools in our sample that failed to make AYP in Vermont are meeting expected targets for their overall populations but failing because of the performance of individual subgroups.[2]

- In Vermont, as in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is much lower. In other words, schools

Fifteen of 18 elementary schools and 17 of 18 middle schools in our sample fail to make AYP in 2008 under **Vermont's** accountability system. This places Vermont at about the middle of the state distribution in terms of the number of schools making AYP. Vermont's proficiency standards are about average compared to other states, but its annual targets are fairly rigorous (roughly 87 percent of grade 3-8 students are expected to be proficient in reading in 2008). Unlike most states, Vermont measures its student performance with a proficiency index, which gives partial credit for students achieving "partial proficiency." In the short term, the index makes it easier for Vermont schools to meet their targets, but the effect of the index diminishes as the targets approach the 100 percent proficiency requirement dictated under NCLB for 2014.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the New England Common Assessment Program (NECAP).

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

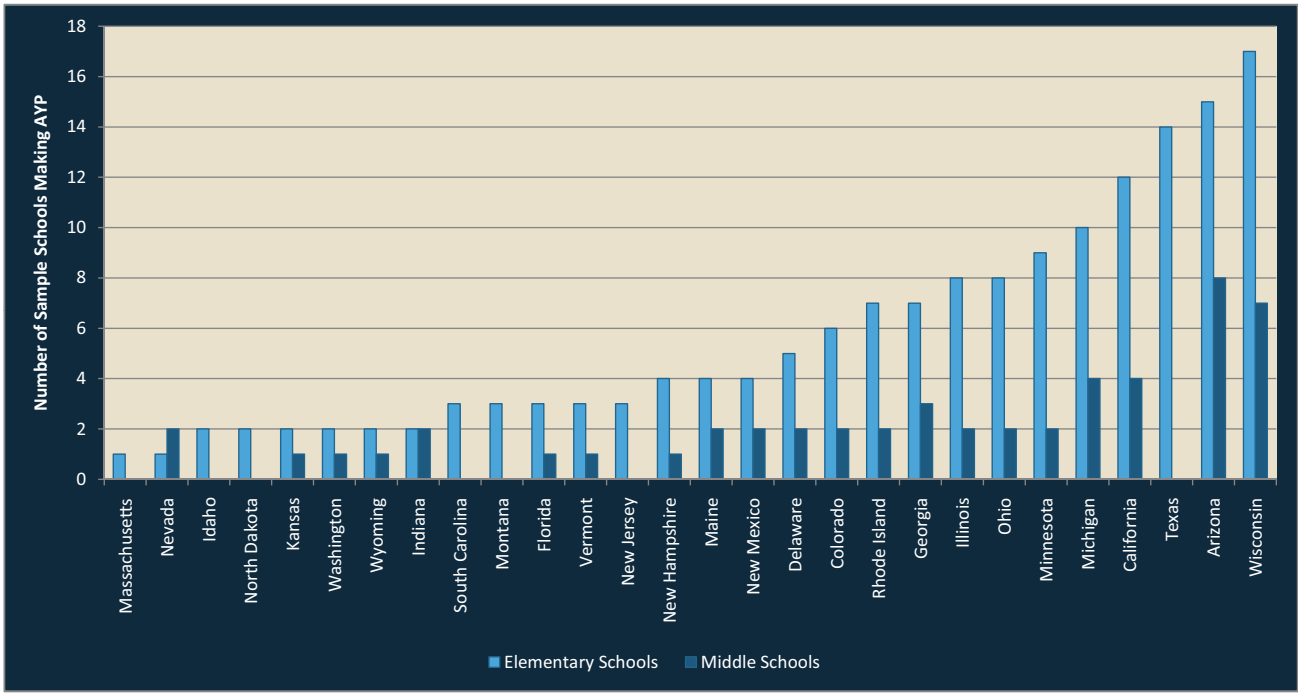

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

with greater diversity and size face greater challenges in making AYP.

- Middle schools have greater difficulty reaching AYP in Vermont than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of whether or not a school will make AYP under the Vermont system is whether it has enough students with disabilities (SWDs)[3] or English language learners to qualify as a separate subgroup. In fact, all schools with limited English proficient (LEP)[4] or SWD subgroups failed to make AYP.

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Vermont's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability

[3] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 New England Common Assessment Program (NECAP), the standardized state test. 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.

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

plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group or as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

Proficiency cut score estimates for the New England Common Assessment Program (NECAP) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Vermont's proficiency cut scores were generally ranked about average compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Vermont test, given their performance on MAP. Student test data and subgroup designations were then used to determine how these 18 elementary and 18 middle schools would have fared under Vermont AYP rules for 2008. In other words, the school data and proficiency cut score esti-

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

[6] We gave all schools in our sample pseudonyms in this report.

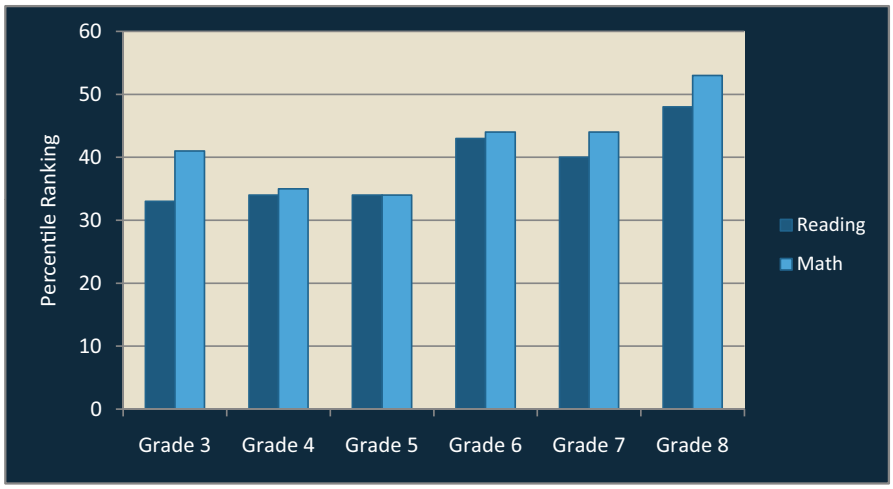

Figure 2. Vermont reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Vermont's cut scores (or proficiency passing scores) for the state's reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of the state's cut scores are below the 55th percentile.

mates are from academic year 2005–2006, but we are applying them against Vermont's 2008 AYP rules.

Table 1 shows the pertinent Vermont AYP rules that were applied to elementary and middle schools in the current study. Vermont's minimum subgroup size is 40, which is comparable to most other states we examined. Most states examined also apply confidence intervals (or margins of error) to their measurements of student proficiency rates. However, Vermont's 99% confidence interval provides schools with greater leniency than the more commonly used 95% confidence interval. This means that while schools are supposed to get 87% of their grade 3-8 students to the proficient level on the state reading test, as well as 87% of the students in each subgroup, applying the confidence interval means that the real target can be lower, particularly with smaller groups.

Unlike most states, Vermont measures its student performance with a proficiency index, which gives partial credit for students achieving "partial proficiency." In the short term, the index makes it easier for Vermont schools to meet their targets, although the effect of the index diminishes as the targets approach the 100% proficiency requirement dictated under NCLB for 2014.[7]

Note that we were unable to examine the impact of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. (Most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. Plus, federal law requires 95% of each school's students—and 95% of the students in each subgroup—to participate in testing.)

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population

[7] In six of the states studied (Massachusetts, Minnesota, Rhode Island, New Hampshire, and Wisconsin, as well as Vermont), an index is used that 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.

Table 1. Vermont AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 40	
	Low-income students: 40	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (index)**	**2008 targets (index)**
READING/LANGUAGE ARTS		
Grade 3	n/a	87.0
Grade 4	n/a	87.0
Grade 5	n/a	87.0
Grade 6	n/a	87.0
Grade 7	n/a	87.0
Grade 8	n/a	87.0
MATH		
Grade 3	n/a	85.4
Grade 4	n/a	85.4
Grade 5	n/a	85.4
Grade 6	n/a	85.4
Grade 7	n/a	85.4
Grade 8	n/a	85.4

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not applicable

and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under Vermont's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Vermont's 2008 AYP rules. **Only three elementary schools made AYP while fifteen failed to make it**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student and positive values indicating above-grade-level performance. All schools making AYP are in the right half of the figure, meaning that they are among the schools which contain the highest average performing students.

Yet among these schools with high average performing students, the only schools actually to make AYP are those with relatively few qualifying subgroups—and thus the fewest targets to meet (since each subgroup has its own separate targets). For example, Wayne Fine Arts, Winchester and Roosevelt made it, but have only four targets each—two in reading and math for their overall populations, and two in reading and math for the only subgroup that exceeds Vermont's minimum "*n* size": white students.

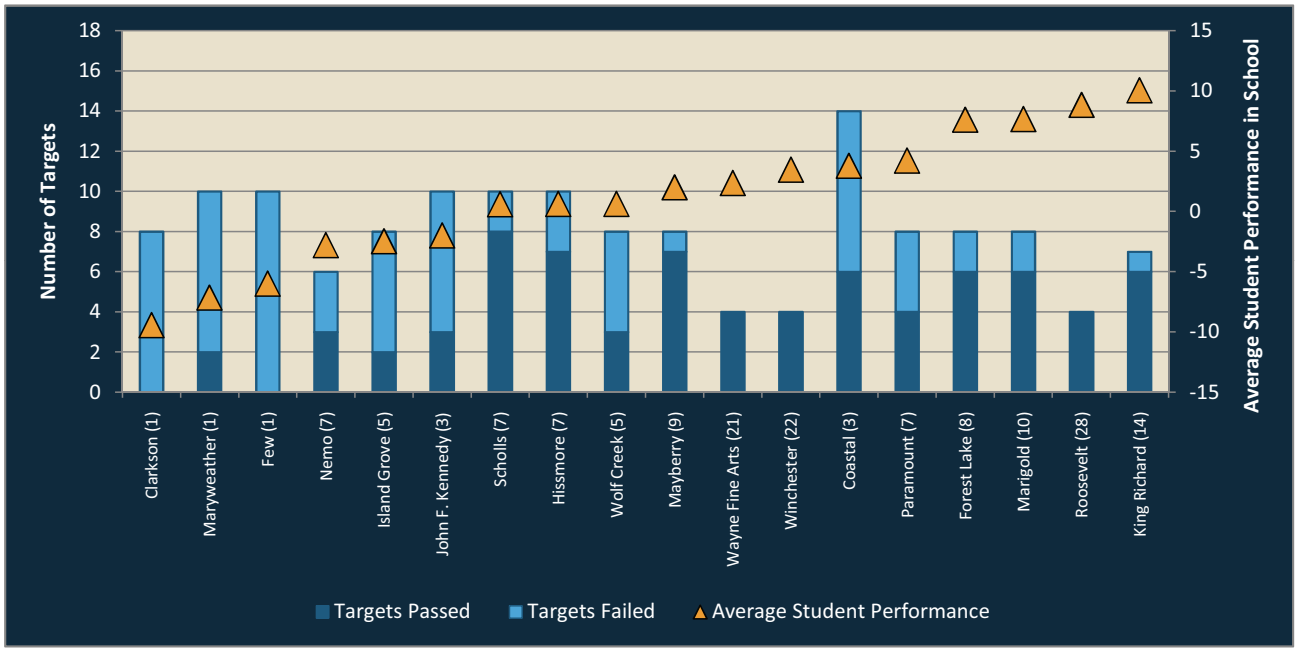

Figure 3. AYP performance of the elementary school sample under Vermont's 2008 AYP rules

Note: This figure shows how each of the elementary schools within the sample fared under the Vermont AYP rules (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Marigold Elementary, for example, meets six of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Vermont AYP rules. **Out of eighteen in our sample, only one middle school made AYP**—Walter Jones—a high-performing school with relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the darker portions of the bars show the actual proficiency rates at each school and the lighter portions of the bars show the degree to which these proficiency rates were increased by applying the confidence interval. The orange lines show the AMOs needed to meet AYP. The figures show that one elementary (JFK) and no middle schools are assisted in meeting their overall math targets by the confidence intervals. However, JFK still failed to make AYP due to the performance of multiple subgroups (see Figure 3).

The effect of the confidence intervals on reading proficiency rates at the elementary and middle school levels is similar (not shown). In reading, two elementary schools (Hissmore and Paramount) and two middle schools (Pogesto and Artemus) were able to meet the overall target with the confidence interval, although we know from Figures 3 and 4 that these schools still failed to meet targets for their subgroups. **In short, applying the confidence interval (even a generous one like the 99% confidence interval used in Vermont) has little or no effect on whether schools meet their overall reading and math targets in Vermont.**[8]

[8] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to *show* how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

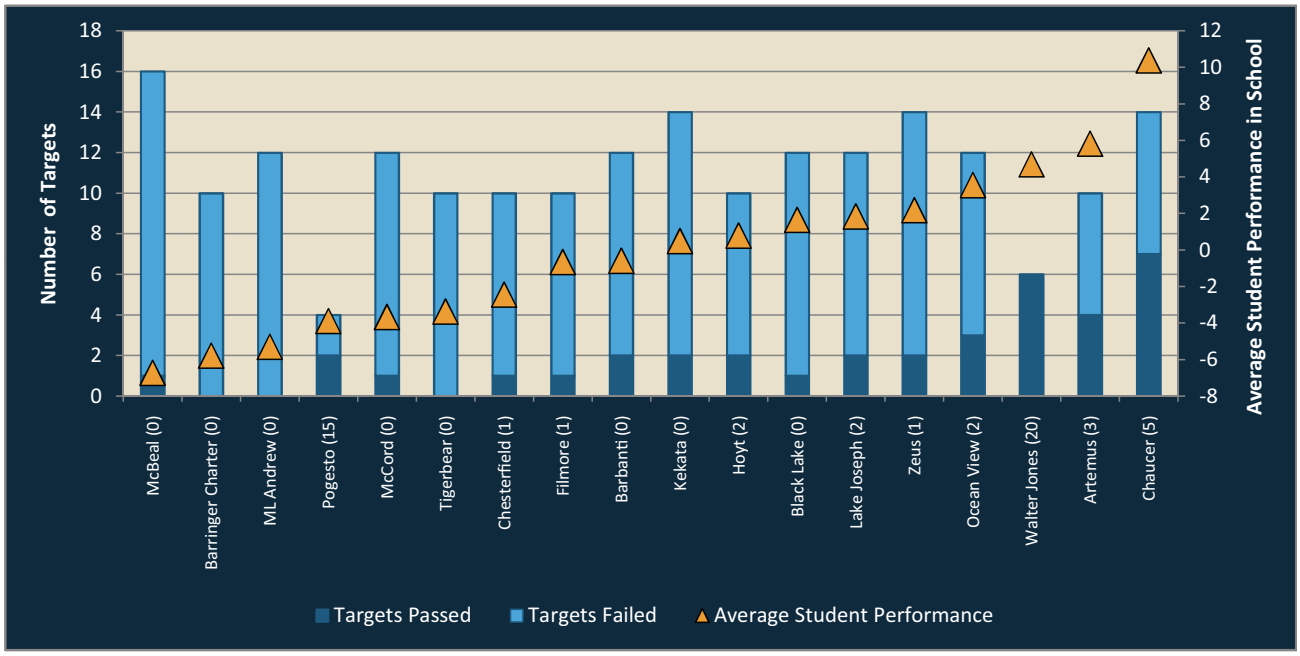

Figure 4. AYP performance of the middle school sample under Vermont's 2008 AYP rules

Note: This figure shows how each of the middle schools within the sample fared under the AYP rules in Vermont (as described in Table 1). The bars show the number of targets that each school had to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Chaucer, for example, meets seven of its fourteen targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles). This is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

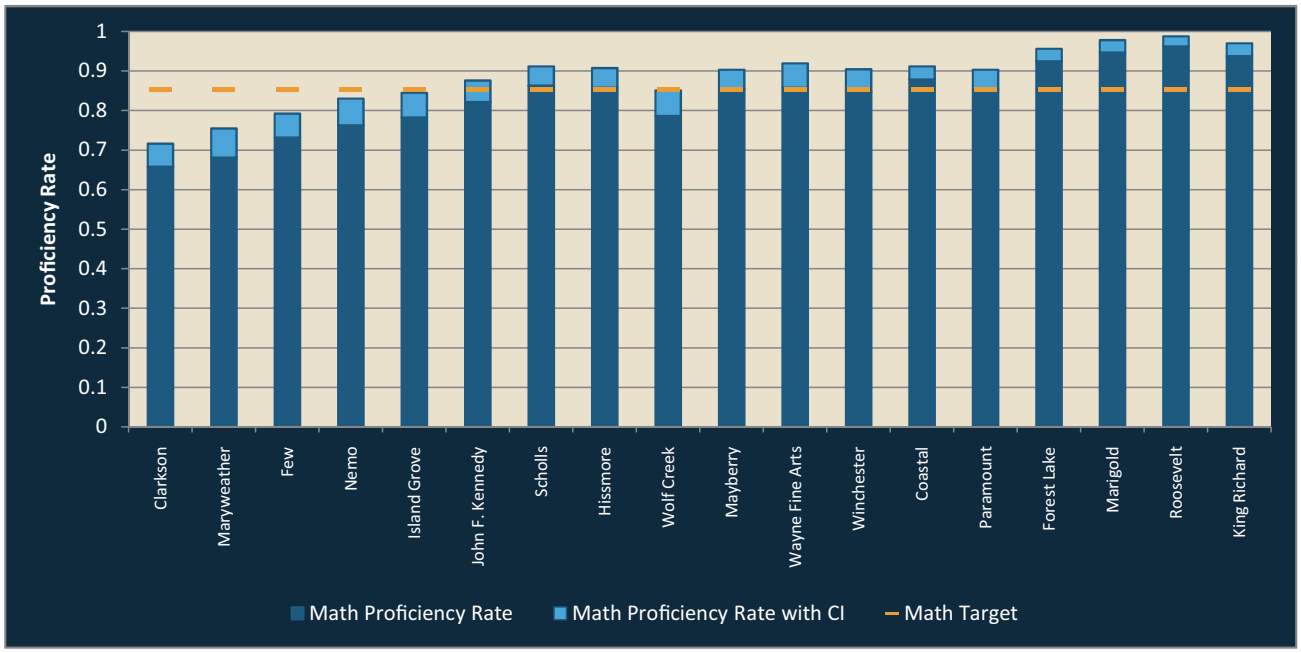

Figure 5. Impact of the confidence interval on elementary school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample elementary schools (JFK) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

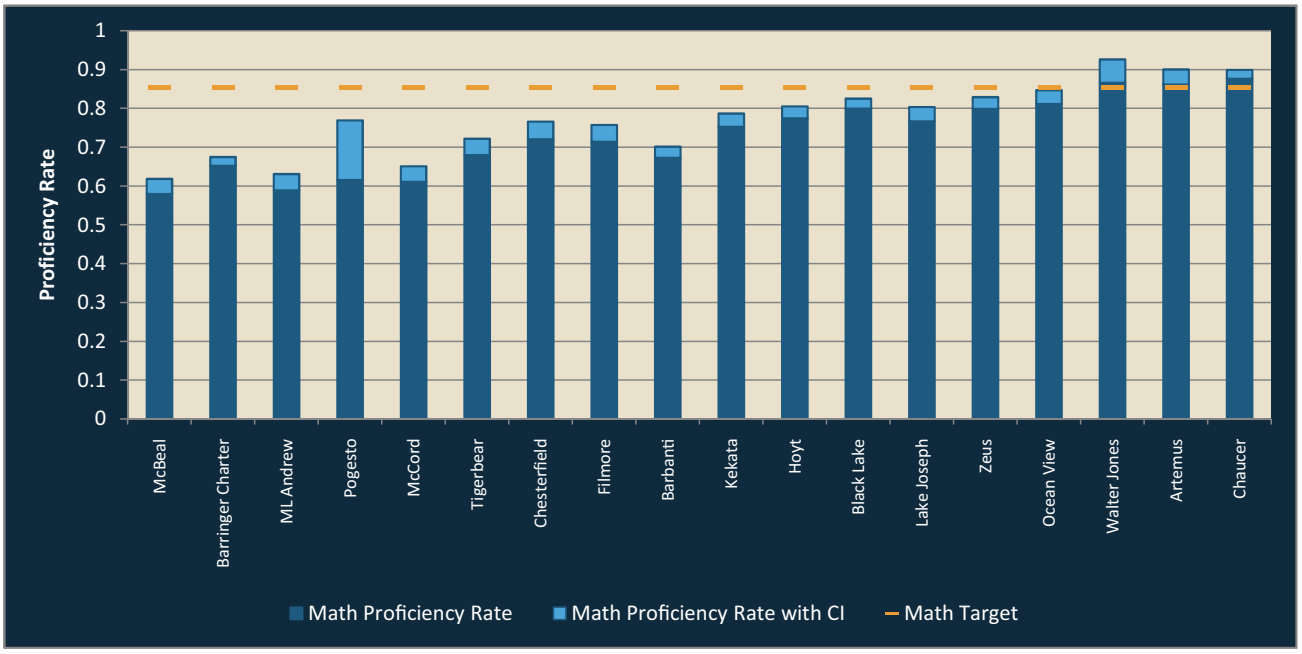

Figure 6. Impact of the confidence interval on middle school math proficiency rates

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample middle schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

Where do schools fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still make AYP when the school has fewer targets to meet, thanks to fewer subgroups. These figures do not, however, indicate which subgroups failed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Vermont rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that

- Four elementary schools failed to meet both their overall reading and math targets.
- Thirteen middle schools failed to meet both their reading and math targets for their overall populations.
- Three elementary schools (Scholls, Forest Lake, and King Richard) failed for their SWD subgroup only.
- One elementary school (Alice Mayberry) met targets for every subgroup except for its low income students.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, the performance of SWDs and LEP students were particularly challenging for Vermont schools. Every single school with enough students to comprise a SWD or LEP subgroup failed to make AYP, in part due to these groups' performances. Traditionally academically disadvantaged subgroups, such as low income and Hispanic students, also had difficulty under Vermont's accountability system, especially at the middle school level.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Vermont AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	65.8%	62.0%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	68.2%	66.4%	N	N			N	N	N	N					N	N			Y	Y	10	2	20%	N	1
Few	73.2%	69.7%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Nemo	76.3%	81.5%	N	Y					N	N									Y	Y	6	3	50%	N	7
Island Grove	78.3%	79.9%	N	N					N	N					N	N			Y	Y	8	2	25%	N	4
JFK	82.2%	78.0%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	3
Scholls	86.3%	81.7%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Hissmore	85.7%	84.3%	Y	Y	N	N			Y	Y	Y	N							Y	Y	10	7	70%	N	7
Wolf Creek	78.7%	81.2%	N	Y					N	N					N	N			Y	Y	8	3	38%	N	5
Alice Mayberry	85.5%	86.2%	Y	Y					Y	N	Y	Y							Y	Y	8	7	88%	N	9
Wayne Fine Arts	85.8%	92.4%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	84.7%	88.3%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	87.9%	83.4%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	3
Paramount	85.3%	84.6%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	7
Forest Lake	92.4%	92.0%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	94.7%	91.2%	Y	Y	Y	N			Y	N									Y	Y	8	6	75%	N	10
Roosevelt	96.2%	96.2%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	93.8%	93.5%	Y	Y	Y	N			Y										Y	Y	7	6	86%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Vermont's NCLB accountability system is, in many respects, behaving similarly to those in other states. For example, among the elementary schools in our sample, Roosevelt, Winchester, and Wayne Fine Arts all made AYP in the greatest number of states—28, 22, and 21, respectively. And these schools all made AYP in Vermont, too. Likewise, the elementary and middle schools that fail to make AYP in the greatest number of states also fail AYP in Vermont. A striking difference between schools that consistently make and don't make AYP, appears to be the number of subgroups for which each is held accountable —and hence, the number of academic targets for which each must demonstrate proficiency.

This is consistent with the patterns shown in Table 6, which compares the schools that did and didn't make AYP on several academic and demographic dimensions. Within the sample, elementary schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller stu-

Table 3. Middle school subgroup performance of sample schools under the 2008 Vermont AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	58.0%	65.0%	N	N	N	N	N	N	N	N	N	N			N	N	N	N	N	Y	16	1	6%	N	0
Barringer Charter	65.3%	73.1%	N	N	N	N			N	N	N	N			N	N					10	0	0%	N	0
ML Andrew	58.9%	71.2%	N	N	N	N			N	N	N	N			N	N			N	N	12	0	0%	N	0
Pogesto	61.6%	74.5%	N	Y															N	Y	4	2	50%	N	15
McCord Charter	61.1%	73.9%	N	N	N	N			N	N	N	N			N	N			N	Y	12	1	8%	N	0
Tigerbear	68.0%	68.8%	N	N	N	N			N	N	N	N							N	N	10	0	0%	N	0
Chesterfield	72.1%	72.9%	N	N	N	N			N	N	N	N							Y	N	10	1	10%	N	1
Filmore	71.4%	78.9%	N	N	N	N			N	N					N	N			N	Y	10	1	10%	N	1
Barbanti	67.3%	74.1%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Kekata	75.3%	76.5%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	0
Hoyt	77.5%	79.3%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	2
Black Lake	80.0%	78.7%	N	N	N	N			N	N	N	N			N	N			Y	N	12	1	8%	N	0
Lake Joseph	76.7%	82.4%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	2
Zeus	79.9%	80.5%	N	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	2	14%	N	1
Ocean View	81.1%	87.3%	N	Y	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	2
Walter Jones	86.6%	88.9%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	86.2%	85.9%	Y	Y	N	N			N	N					N	N			Y	Y	10	4	40%	N	3
Chaucer	87.7%	91.0%	Y	Y	N	N	N	N	N	N			Y	Y	N	Y			Y	Y	14	7	50%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

dent populations, fewer subgroups (and thus fewer targets to meet), and lower percentages of low-income students. Similarly, middle schools that make AYP have slightly higher performing students, on average, than middle schools that failed to make it, but have dramatically smaller total enrollments, smaller nonwhite populations, and fewer subgroups (and thus targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would have fared under Vermont's AYP rules and annual measurable objectives for 2008. We found that only 3 elementary schools and 1 middle school—4 in all from a sample of 36—would have made AYP in Vermont. Looking across the 28 state accountability systems examined in this study, this puts Vermont at about the middle of the distribution in terms of the number of elementary sample schools making AYP (as shown in Figure 1).

Because the overriding goal of NCLB is to eliminate educational disparities within and across states, it's impor-

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Vermont AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	6	8
Students with limited English proficiency	4	4	4
Low-income students	15	8	11
African-American students	5	1	3
Asian/Pacific Islander students	0	0	0
Hispanic students	7	7	7
American Indian/Alaska Native students	0	0	0
White students	16	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Vermont AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	7	7	7
Low-income students	17	16	16
African-American students	10	10	10
Asian/Pacific Islander students	1	0	0
Hispanic students	13	13	12
American Indian/Alaska Native students	1	1	1
White students	17	6	4

tant to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Vermont's No Child Left Behind accountability system is working exactly as Congress intended: identifying as needing attention schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Some of the sample schools made AYP in Vermont for their student populations as a whole. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, etc. has made those students visible. That is surely a good thing.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has

Table 6. Comparisons between schools that did and didn't make AYP in Vermont, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	3	15	1	17
Average student body size	225	321	165	900
Average % low income	16	52	38	45
Average % nonwhite	27	44	33	45
Average performance†	4.89	0.49	4.69	-0.33
Average % growth‡	113	115	111	97
Average number of targets to meet	4	9	6	11

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[9] doesn't the massive failure of middle school students to meet Vermont's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Is it "fair" that, in Vermont and in a handful of other states, students are awarded "partial" credit even though they do not achieve proficiency? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency

[9] See footnote 3.

ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Washington

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all of their students achieve mastery in reading and math, 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 goals.

This report examines Washington's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either making AYP—or not making AYP. It also gauges how tough Washington's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Washington's system as well as under the systems 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Washington's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **16 of 18 elementary schools** and **17 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Washington's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students.

- It's also partly explained by Washington's somewhat smaller minimum *n* size for its race/ethnicity and low income subgroups, which means more of these students are held separately accountable in Washington than they might be in other states. In addition, Washington has above average proficiency standards, especially at the middle school level, and relatively high annual targets, especially in grades 3-5 reading. Both these factors potentially hinder a school's chance of making AYP in Washington.

- Looking across the 28 state accountability systems examined in the study, **we find that the number of elementary schools that made AYP in Washington is exceeded in 20 other sample states (Washington ties with 5 other states that each has 2 schools that made AYP). In addition, Washington is one of 6 states with a single middle school making AYP (See Figure 1).**

- Most of the schools in our sample that failed to make AYP in Washington are meeting expected targets for

Only two elementary schools and one middle school in our sample make AYP in 2008 under **Washington's** accountability system. One of the main reasons is that the state has a relatively small subgroup size for its minority and low-income students. This means that schools in Washington may have more accountable minority and low-income subgroups than would similar schools in other states. In addition, Washington has above average proficiency standards, especially at the middle school level, and relatively high annual targets, especially in grades 3-5 reading. Even though Washington's 99 percent confidence interval (i.e., statistical margin of error) provides schools with greater leniency than the more commonly used 95 percent confidence interval, these other factors make Washington among the most restrictive states in terms of the number of schools making AYP.

[1] A cut score is the minimum score a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Washington Assessment of Student Learning (WASL).

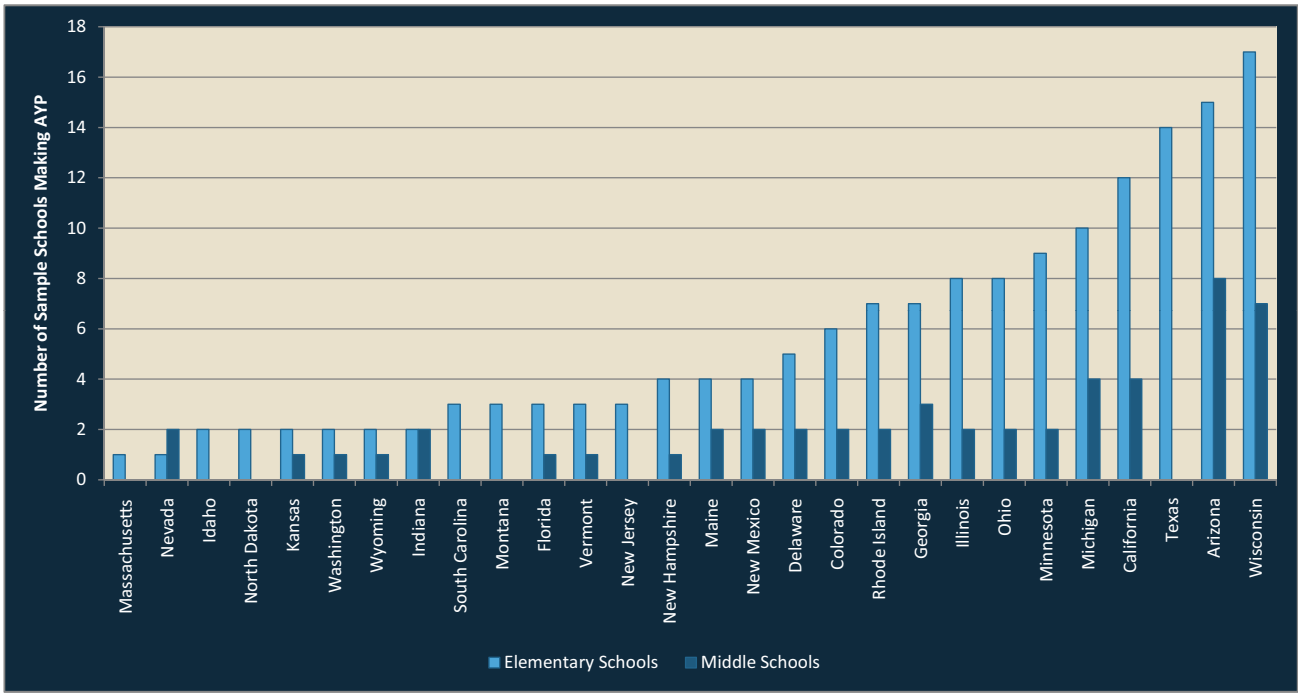

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

their overall populations[2] but failing because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.

- In Washington, as in most states, schools with fewer subgroups attained AYP more easily than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As is the case in other states, middle schools have greater difficulty reaching AYP in Washington than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of whether or not a school would make AYP under Washington's system is whether it has enough English language learners to qualify as a separate subgroup. Every single school with a limited English proficient (LEP)[3] subgroup failed to make AYP, in part because these students did not meet the state's targets in reading and math. Likewise, every school with enough qualifying SWDs failed to make AYP.[4]

Introduction

The Proficiency Illusion (Cronin, et, al. 2007a) linked student performance on Washington's tests and those of 25

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

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

[4] SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP and SWDs may be slightly more negative than would be seen under real world conditions. This is mostly due to the differences in testing practices between how LEP students and students with disabilities are treated in the Washington Assessment of Student Learning (WASL) state assessment and in the NWEA's Measures of Academic Progress (MAP), the assessment used in this study. Specifically, the U.S. Department of Education has issued NCLB guidelines permitting schools to exclude small percentages of LEP or disabled students from taking state tests, or providing them alternate assessments. In the current study, however, no valid MAP scores were omitted from consideration.

other state tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[5] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error–to try to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[6]

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

[6] We gave all schools in our sample pseudonyms in this report.

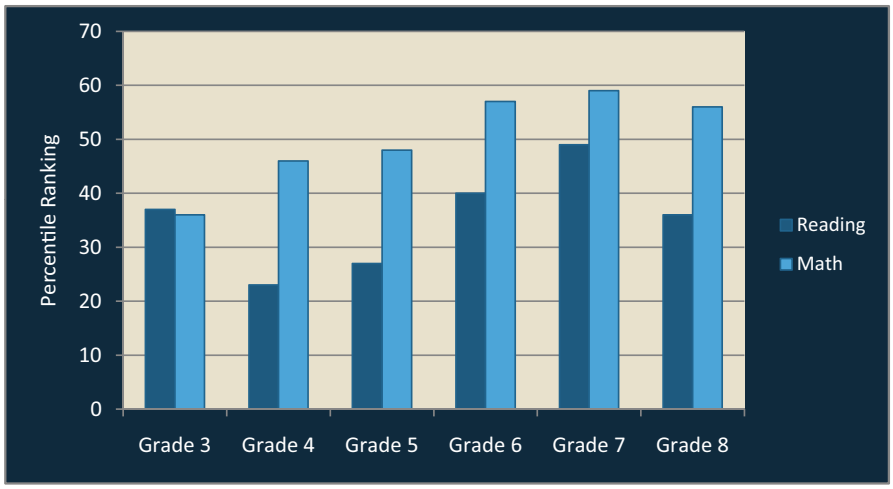

Figure 2. Washington reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Washington's cut scores (proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades 3 through 8. Higher percentile ranks are more difficult to achieve. All of Washington's cut scores are below the 60th percentile.

Proficiency cut score estimates for the Washington Assessment of Student Learning (WASL) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Washington's proficiency cut scores generally ranked above the average in difficulty, compared with the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Washington test, given their performance on MAP. Student test data and subgroup designations are then used to determine how these 18 elementary and 18 middle schools would have fared under Washington AYP rules for 2008. (In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Washington's 2008 AYP rules.)

Table 1 shows the pertinent Washington AYP rules that were applied to elementary and middle schools in the current study. Washington's minimum *n* sizes, unlike other states, vary according to subgroup. The subgroup size is 30 for the race/ethnicity and low-income subgroups and 40 for SWDs and students with limited English proficiency. For schools with 4000 or more students, the subgroup minimums for these last two groups is 1% of the school population. A subgroup size of 40 is typical, compared to other states in the study, but 30 is a bit lower.[7] This means that schools in Washington may have more accountable subgroups than would similar schools in other states.

Most states examined also apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates. However, Washington's 99% confidence interval provides schools with greater leniency than the more commonly used 95% confidence interval. This means even though the AMO might require a school to attain, for instance, 76.1% reading proficiency among its grade 3 students, and 76.1% reading proficiency among its grade 3 students in each subgroup, the real target can be lower, particularly with smaller groups.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail as long as it reduces by at least 10% the number of non-proficient students within any failing subgroup, relative to the previous year's performance. Because we had access to only a single year's data (2005–2006), we were not able to include this in our analysis. As a result, it's possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

[7] School size and *n* size, however, are related (e.g., it makes sense for small schools to have small *n* sizes).

Table 1. Washington AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 40, or 1% if school population > 4000	
	Low-income students: 30	
	LEP students: 40, or 1% if school population > 4000	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	76.1
Grade 4	52.2	76.1
Grade 5	n/a	76.1
Grade 6	n/a	65.1
Grade 7	30.1	65.1
Grade 8	n/a	65.1
MATH		
Grade 3	n/a	64.9
Grade 4	29.7	64.9
Grade 5	n/a	64.9
Grade 6	n/a	58.7
Grade 7	17.3	58.7
Grade 8	n/a	58.7

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a = not available

Furthermore, attendance and test participation rates are beyond the scope of the study. Most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. Plus, federal law requires 95% of each school's students, and 95% of students in each subgroup, to participate in testing.

So, to reiterate, AYP decisions in the current study are modeled solely on test performance data for a single year. For each school, overall reading and math proficiency rates are calculated (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the annual measurable objectives. A school is deemed to have made AYP if the overall student body and all its qualifying subgroups met or exceeded its annual measurable objectives. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare Under Washington's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Washington's 2008 AYP rules. **Only 2 elementary schools out of 16 made AYP**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance, and positive values indicating above-grade-level performance. The schools that made AYP are in the right half of the figure,

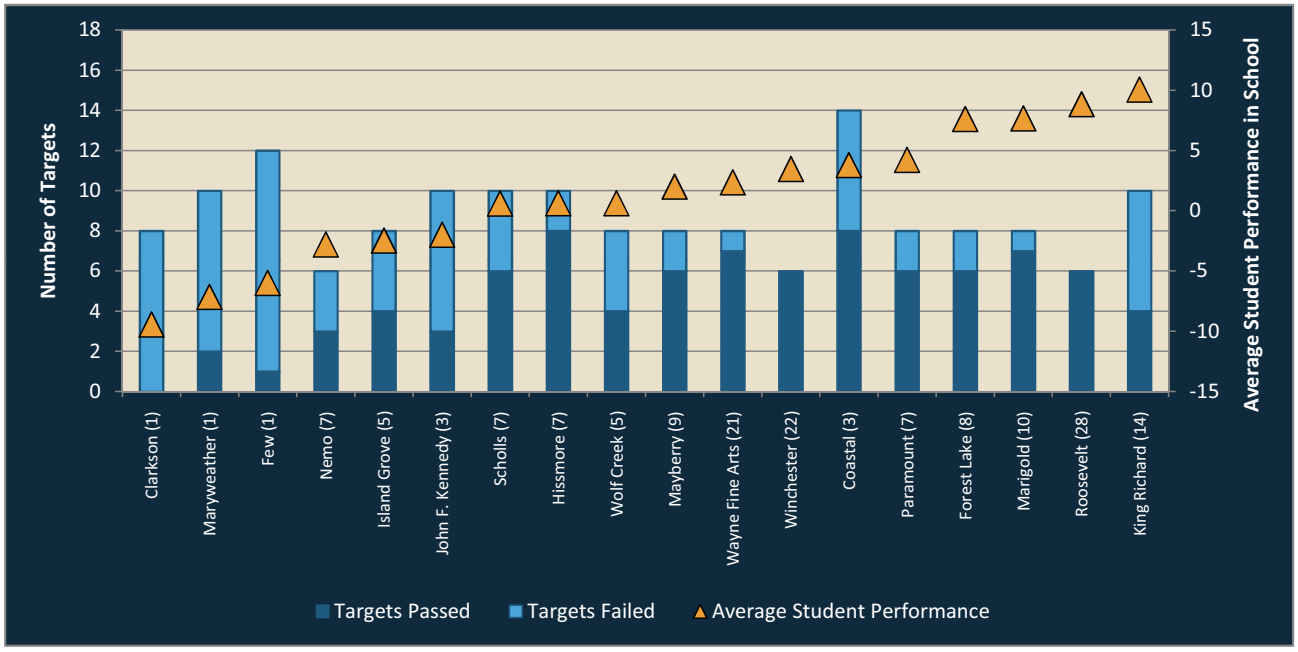

Figure 3. AYP performance of the elementary school sample under Washington's 2008 AYP rules

Note: This figure indicates how each of the elementary schools within the sample fared under Washington's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Marigold Elementary, for example, met seven of its eight targets, but because it did not meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

meaning the higher performing students were found at these schools.

Yet almost without regard to average student performance, the only schools actually to make AYP were those with relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has separate targets). For example, Winchester and Roosevelt made AYP, but have only six targets each. Each had to meet two targets in reading and math for their overall student population, two more targets for their white subgroup, and two more targets for an additional subgroup—Hispanic for Winchester and low income for Roosevelt.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Washington AYP rules. **Out of eighteen in our sample, only one made AYP**—a high-performance school (Walter Jones), that has relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates were increased by the application of the confidence interval. The orange lines show the annual measurable objective needed to meet AYP. These figures show that three of the sample elementary schools (Island Grove, JFK, and Wolf Creek) and one of the middle schools (Kekata) is assisted by the confidence intervals (note how the orange line falls within the light blue bar). However, we know from Figures 3 and 4 that all of these schools still fail to make AYP because of low subgroup performance.

The effect of confidence intervals on reading proficiency rates for elementary and middle schools is much the same (not shown). In reading, four elementary schools

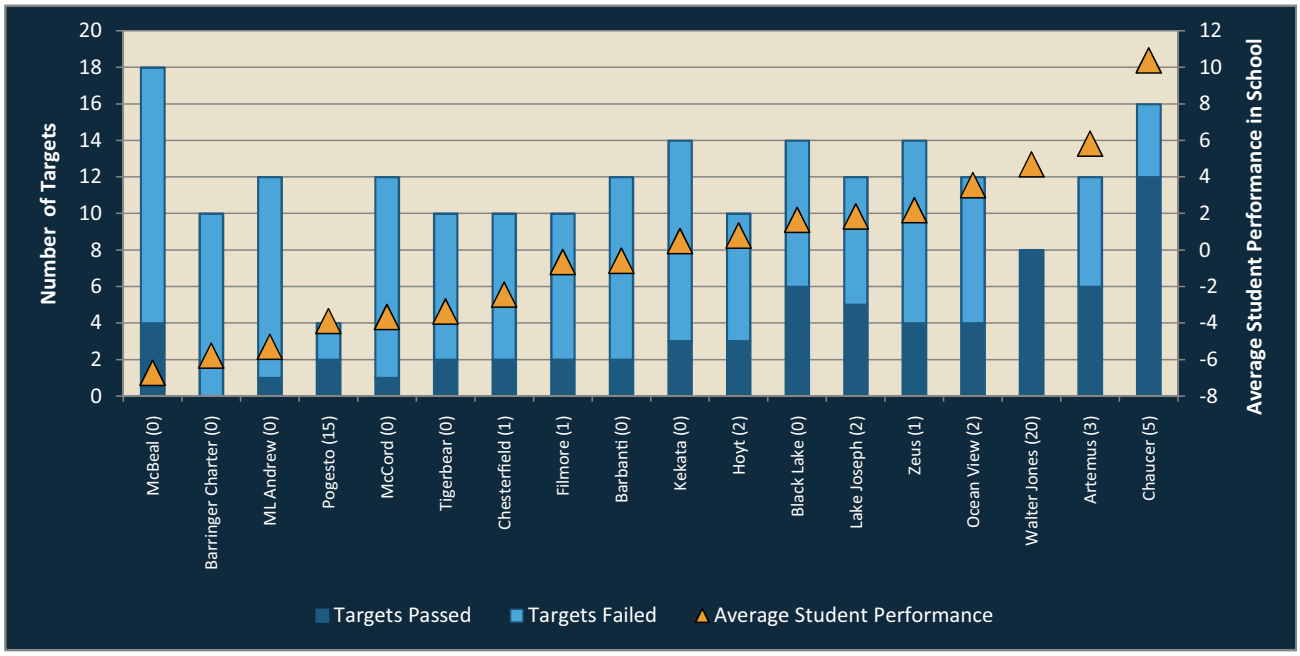

Figure 4. AYP performance of the middle school sample under Washington's 2008 AYP rules

Note: This figure indicates how each of the middle schools within the sample would have fared under Washington's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet in order to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup didn't make AYP, so any light blue means the school failed. Pogesto, for example, meets two of its four targets, but because it did not meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

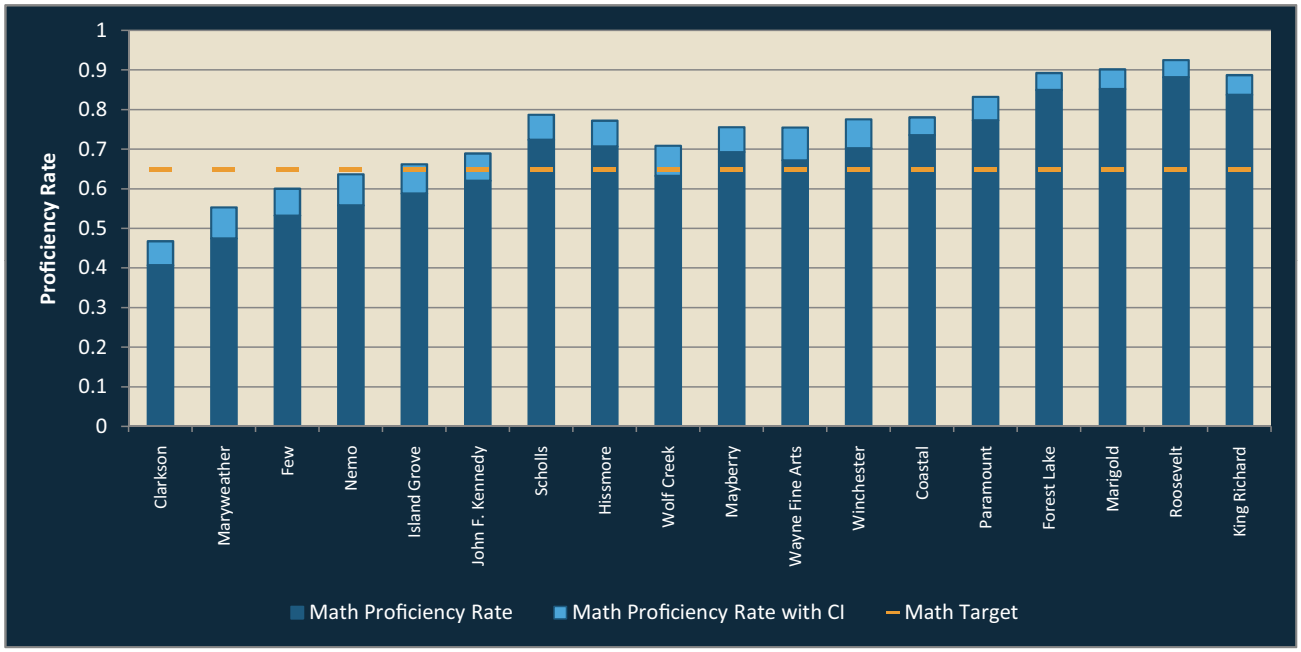

Figure 5. Impact of the confidence interval on elementary school mathematics proficiency rates under Washington's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that three of the sample elementary schools (Island Grove, JFK, and Wolf Creek) were assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

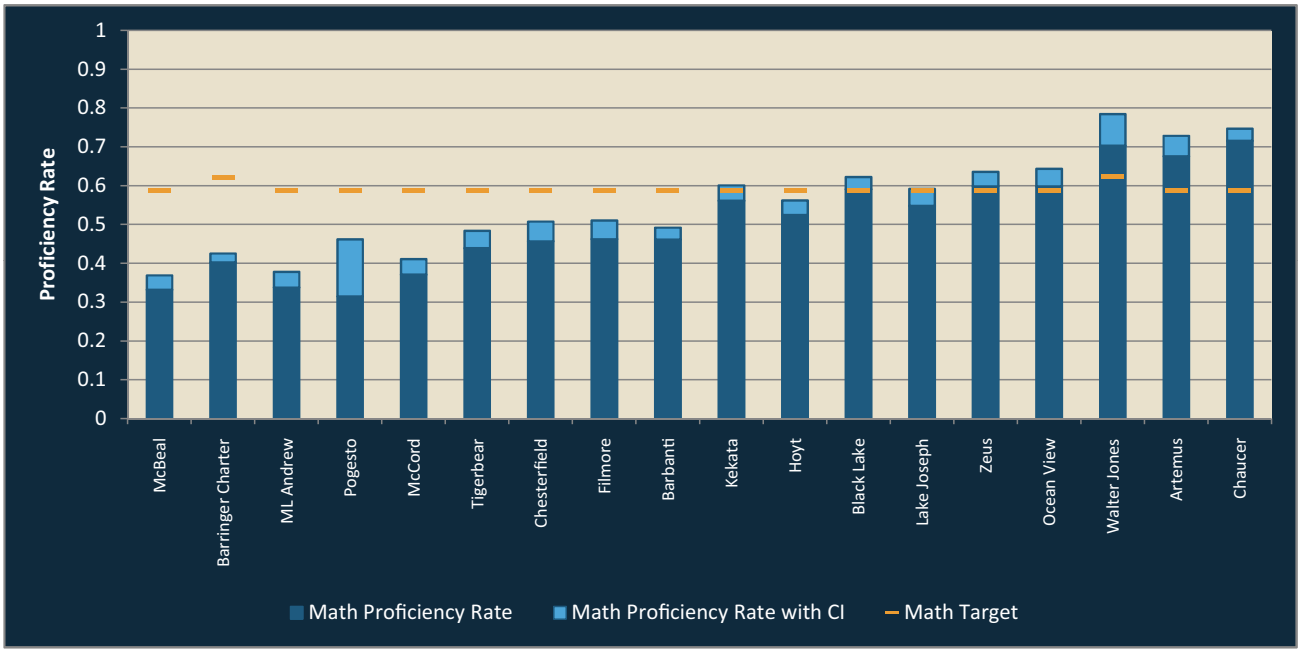

Figure 6. Impact of the confidence interval on middle school mathematics proficiency rates under Washington's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools (Kekata) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

(Nemo, Island Grove, Scholls and Wolf Creek) and two middle schools (Pogesto and Black Lake) are able to meet the overall target with the confidence interval, but still fail to meet their targets for all required subgroups. **In short, the application of the confidence interval has little effect on final AYP decisions for the sample schools in Washington, even though it does help some schools to meet their overall targets.**[8]

Where do schools fail?

Figures 3 and 4 illustrate how schools with low or middling performance can still pass AYP when the school has few targets to meet, thanks to fewer subgroups. However, these figures do not indicate which subgroups failed or passed in which school. Information on individual subgroup performance appears in Tables 2 and 3 for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. While all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP purposes include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American (AA), Asian/Pacific Islander (Asian), Hispanic/Latino (Hispanic), American Indian/Alaska Native (AI/AN), and white. Tables 2 and 3 also show whether a school made AYP under the Washington rules, and the total number of states within the study in which the school made AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Three elementary schools (Clarkson, Maryweather,

[8] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary subgroup performance of sample schools under the 2008 Washington AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	40.7%	43.7%	N	N			N	N	N	N					N	N					8	0	0%	N	1
Maryweather	47.5%	55.7%	N	N			N	N	N	N					N	N			Y	Y	10	2	20%	N	1
Few	53.3%	56.0%	N	N	N	N	N	N	N	N					N	N			Y	N	12	1	8%	N	1
Nemo	55.8%	68.8%	N	Y					N	N									Y	Y	6	3	50%	N	7
Island Grove	58.8%	71.2%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	4
JFK	62.1%	63.8%	Y	N	N	N			N	N	N	N							Y	Y	10	3	30%	N	3
Scholls	72.4%	72.1%	Y	Y	N	N			Y	Y	N	N							Y	Y	10	6	60%	N	7
Hissmore	70.7%	75.2%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	63.3%	70.3%	Y	Y					N	N					N	N			Y	Y	8	4	50%	N	5
Alice Mayberry	69.3%	75.4%	Y	Y					Y	N	Y	N							Y	Y	8	6	75%	N	9
Wayne Fine Arts	67.2%	85.1%	Y	Y					Y	Y	N	Y							Y	Y	8	7	88%	N	21
Winchester	70.3%	80.6%	Y	Y											Y	Y			Y	Y	6	6	100%	Y	22
Coastal	73.6%	80.3%	Y	Y	N	N	N	N	Y	Y	N	N			Y	Y			Y	Y	14	8	57%	N	3
Paramount	77.3%	77.2%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	85.0%	86.0%	Y	Y	N	N			Y	Y									Y	Y	8	6	75%	N	8
Marigold	85.3%	89.9%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	10
Roosevelt	88.2%	92.5%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	83.8%	91.5%	Y	Y	N	N			N	N					N	N			Y	Y	10	4	40%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

and Few) failed to meet both the reading targets and math targets for their overall school population.

- Three elementary schools (Hissmore, Marigold, and Forest Lake) met all targets except for their SWDs, and one school (Wayne Fine Arts) met all required targets except for its African American subgroup.

- Eight of the sample middle schools failed to meet both their reading and math targets for their overall population.

- None of the schools with qualifying SWD subgroups and LEP subgroups met AYP.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, the performance of SWDs and students with LEP are particularly challenging for Washington schools. Almost every single school with a large enough population of students in these groups to exceed the minimum *n* size failed to meet their subgroup targets. Nearly all of the traditionally academically-disadvantaged subgroups (e.g., low income and African American) also struggle under Washington's accountability system.

Table 3. Middle school subgroup performance of sample schools under the 2008 Washington AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	33.2%	47.5%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	40.2%	56.1%	N	N	N	N			N	N	N	N			N	N					10	0	0%	N	0
ML Andrew	33.8%	51.1%	N	N	N	N			N	N	N	N			N	N			N	Y	12	1	8%	N	0
Pogesto	31.5%	53.7%	N	Y															N	Y	4	2	50%	N	15
McCord Charter	37.1%	54.8%	N	N	N	N			N	N	N	N			N	N			N	Y	12	1	8%	N	0
Tigerbear	43.9%	51.0%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	0
Chesterfield	45.7%	50.1%	N	N	N	N			N	N	N	N							Y	Y	10	2	20%	N	1
Filmore	46.2%	60.2%	N	N	N	N			N	N					N	N			Y	Y	10	2	20%	N	1
Barbanti	46.2%	55.9%	N	N	N	N	N	N	N	N					N	N			Y	Y	12	2	17%	N	0
Kekata	56.1%	60.1%	Y	N	N	N	N	N	N	N	N	N			N	N			Y	Y	14	3	21%	N	0
Hoyt	52.5%	62.4%	N	Y	N	N			N	N	N	N							Y	Y	10	3	30%	N	2
Black Lake	59.1%	64.1%	Y	Y	N	N			N	N	N	N	Y	Y	N	N			Y	Y	14	6	43%	N	0
Lake Joseph	54.8%	67.3%	Y	Y	N	N	N	N	N	Y					N	N			Y	Y	12	5	42%	N	2
Zeus	59.9%	66.2%	Y	Y	N	N	N	N	N	N	N	N			N	N			Y	Y	14	4	29%	N	1
Ocean View	59.8%	75.0%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	70.3%	81.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	67.6%	73.8%	Y	Y	N	N			N	N			Y	Y	N	N			Y	Y	12	6	50%	N	3
Chaucer	71.6%	82.3%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Washington's NCLB accountability system is, in some respects, behaving like those in other states. For example, Roosevelt and Winchester are among the schools that make AYP in the greatest number of states—28 and 22, respectively. And these schools make AYP in Washington, too. Likewise, the elementary and middle schools that fail to make AYP in the greatest number of states also fail to make AYP in Washington.

But Washington is home to at least one anomaly. Consider Wayne Fine Arts Elementary (see Figure 3). It failed to make AYP in Washington, but makes AYP in 21 other states in our sample. Examining Table 2, one can see that Wayne Fine Arts failed to meet the minimum numbers for its LEP or SWD subgroups, which create difficulty for so many other schools within the sample. It did, however, miss the math target for its African American subgroup, possibly because of Washington's harder than average proficiency standards.

The differences between schools that did and didn't

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Washington AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	8	7	8
Students with limited English proficiency	4	4	4
Low-income students	17	8	10
African-American students	6	4	4
Asian/Pacific Islander students	0	0	0
Hispanic students	9	6	7
American Indian/Alaska Native students	0	0	0
White students	17	0	1

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Washington AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	7	7	7
Low-income students	17	15	14
African-American students	11	10	10
Asian/Pacific Islander students	4	0	0
Hispanic students	14	12	12
American Indian/Alaska Native students	1	1	1
White students	17	3	0

make AYP under Washington's accountability system can be seen in Table 6, which compares them on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, fewer subgroups (and thus fewer targets to meet), and much lower percentages of low income students.

The picture for middle schools is similar. The one middle school that made AYP had slightly higher performing students, on average, than middle schools that didn't, as well as a drastically smaller enrollment, a smaller non-white population, and fewer subgroups (and thus targets to meet).

Table 6. Comparisons between schools that did and didn't make AYP in Washington, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	2	16	1	17
Average student body size	225	315	165	900
Average % low income	13	50	38	45
Average % nonwhite	25	43	33	45
Average performance†	6.16	0.61	4.69	-0.33
Average % growth‡	121	114	111	97
Average number of targets to meet	6	9	8	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Washington's AYP rules and annual measurable objectives for 2008. We found that only two elementary schools and one middle school—three in all from a total of 36—would have made AYP in Washington. Looking across the 28 state accountability systems examined in the study, this puts Washington at the low end of the distribution in terms of the number of schools making AYP (as shown in Figure 1). This high failure rate is partly explained by Washington's somewhat smaller minimum *n* size for its race/ethnicity and low income subgroups, which means more of these students are held separately accountable in Washington than they might be in other states. In addition, Washington has above average proficiency standards, especially at the middle school level, and relatively high annual targets, especially in reading for grades 3 through 5. All of these factors potentially hinder a school's chance of making AYP in The Evergreen State.

The overriding goal of the No Child Left Behind act (NCLB) is to eliminate educational disparities within and across states; it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Washington's No Child Left Behind accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income or Hispanic students. Many of the sample schools met the Washington math and reading targets for their student populations as a whole (i.e., without considering subgroup results). In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, etc. has made those students visible. That is surely a good thing.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are

more generous under the current state assessment system,[9] doesn't the failure of these students to meet Washington's targets (especially at the middle school level where more of them qualify) indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for LEP students and SWDs, as for other students, but when so few schools are able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB reauthorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

[9] See footnote 4.

Wisconsin

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 towards these goals.

This report examines Wisconsin's NCLB accountability system—particularly how its various rules, criteria, and practices result in schools either "making AYP"—or not making AYP. It also gauges how tough Wisconsin's system is compared with other states. For this study, we selected 36 schools from around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Wisconsin's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Wisconsin's AYP rules.

Here are some key findings:

- We estimate that **just 1 of 18 elementary schools** and **11 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Wisconsin's accountability system.
- **Looking across the 28 state accountability systems examined in the study, we find that Wisconsin has the greatest number of elementary schools making AYP in our sample**. In addition, at seven, Wisconsin has the second highest number of middle schools making AYP in the sample (only Arizona has more) (See Figure 1).
- **The high number of schools making AYP in Wisconsin is likely due to the fact that Wisconsin's proficiency standards are extremely easy compared to other states, plus it uses a proficiency index, which means it gives "partial credit" to students performing below proficient.**
- The few schools in our sample that fail to make AYP in Wisconsin are meeting expected targets for their overall populations[2] but failing because of the performance of individual subgroups, particularly students

More schools make AYP in 2008 under **Wisconsin's** accountability system than in any other state in our sample. This is likely due to the fact that Wisconsin's proficiency standards (or cut scores) are relatively easy compared to other states (all of them are below the 30th percentile). Second, Wisconsin's minimum subgroup size for students with disabilities is 50, which is a bit larger than most other states (the size for their other subgroups is comparable to other states'). This means that Wisconsin schools must have more students with disabilities in order for that group to be held separately accountable. Third, Wisconsin's 99 percent confidence interval provides schools with greater leniency than the more commonly used 95 percent confidence interval. Last, unlike most states, Wisconsin measures its student performance with a proficiency index, which gives partial credit for students achieving "partial proficiency." All of these factors work together so that 17 out of 18 elementary schools make AYP in Wisconsin, more than any other state in the study.

[1] A cut score is the minimum score on a student must receive on NWEA's Measures of Academic Progress (MAP) that is equivalent to performing proficient on the Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT).

[2] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they are simply not treated as their own subgroup.

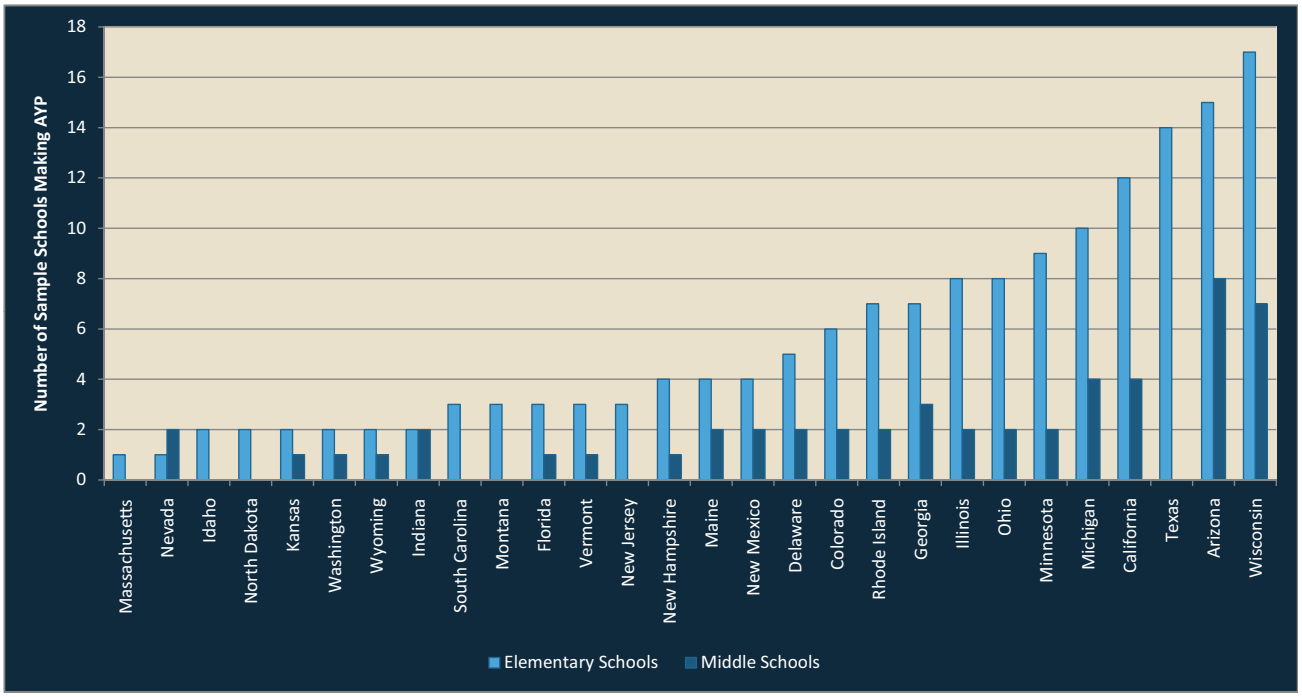

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

with disabilities (SWDs)[3] and students with limited English proficiency (LEP).[4]

- Two sample schools that failed to make AYP in any other state made AYP in Wisconsin. Again, this is likely due to the fact that Wisconsin's proficiency standards are relatively easy compared to other states, along with the fact that these two schools had fewer accountable subgroups.

- In Wisconsin, as is the case in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- As in other states, middle schools have greater difficulty reaching AYP in Wisconsin than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is any lower than in the elementary schools.

- A strong predictor of a school making AYP under Wisconsin's system is whether it has enough SWDs to qualify as a separate subgroup. Almost all schools with qualifying subgroups in this category failed to meet their AYP targets, particularly at the middle school level.[5] **Ironically, Wisconsin has one of the largest minimum *n* sizes for SWDs in our sample; still, when enough SWDs exist to comprise a subgroup, they do not perform well.**

[3] SWDs are defined as those students following individualized education plans.

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

[5] It should be noted that our subgroup findings for Limited English Proficient (LEP) and students with disabilities (SWDs) may be slightly more negative than would be seen under real world conditions. This is mostly due to the differences in testing practices between how LEP students and SWDs are treated in the Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT) state assessment and NWEA's Measures of Academic Progress (MAP), the assessment used in this study. Specifically, the U.S. Department of Education has issued NCLB guidelines permitting schools to exclude small percentages of LEP or disabled students from taking state tests, or providing them alternate assessments. In the current study, however, no valid MAP scores were omitted from consideration.

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Wisconsin's tests and those of 25 other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013–2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. These AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than 10 pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as ten youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Arizona or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008).

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as students with limited English proficiency (LEP), among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-

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

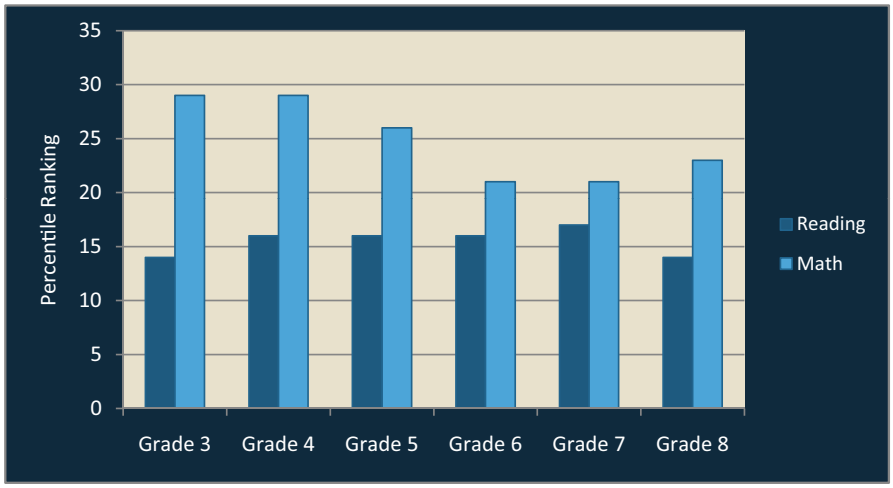

Figure 2. Wisconsin reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Wisconsin's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Wisconsin's math cut scores are below the 30th percentile and all its reading cut scores are below the 18th percentile.

price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

Proficiency cut score estimates for the Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT) are taken from *The Proficiency Illusion* (as shown in Figure 2), which found that Wisconsin's definitions of proficiency were generally below average compared with the standards set by the other 25 states in that study (especially in reading). These cut scores were used to estimate whether students would have scored as proficient or better on the Wisconsin test, given their performance on MAP. Student test data and subgroup designations are then used to determine how these 18 elementary and 18 middle schools would have fared under Wisconsin AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Wisconsin's 2008 AYP rules.

Table 1 shows the pertinent Wisconsin AYP rules that were applied to elementary and middle schools in this study. Wisconsin's minimum subgroup size for most subgroups is 40, which is comparable to most other states we examined.[8] However, for students with disabilities (SWDs) the minimum is 50, which is a bit larger than most other states.

Furthermore, although most states examined in the study apply confidence intervals (or margins of statistical error) to their measurements of student proficiency rates, Wisconsin's 99% confidence interval gives schools greater leniency than the more commonly used 95% confidence interval used by most other states. So, for instance, although schools are supposed to get 74% of their grade 3-8 students (as well as 74% of their grade 3-8 students in each subgroup) to the proficient level on the state reading test, applying the confidence interval means that the real target can actually be lower, particularly with smaller groups.[9]

Unlike most states, Wisconsin measures its student performance with a proficiency index, which gives partial credit for students achieving "partial proficiency." In the short term, the index makes it easier for Wisconsin

[7] We gave all schools in our sample pseudonyms in this report.

[8] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

[9] We also conducted an analysis to show the effect of confidence intervals on the reading and math proficiency rates for elementary and middle schools. We describe those results later in the report.

Table 1. Wisconsin AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 40	
	SWDs: 50	
	Low-income students: 40	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 99% CI used	
AMOs	**Baseline proficiency levels as of 2002 (index)**	**2008 targets (index)**
READING/LANGUAGE ARTS		
Grade 3	61	74
Grade 4	61	74
Grade 5	61	74
Grade 6	61	74
Grade 7	61	74
Grade 8	61	74
MATH		
Grade 3	37	58
Grade 4	37	58
Grade 5	37	58
Grade 6	37	58
Grade 7	37	58
Grade 8	37	58

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives

schools to meet their targets, although the effect of the index diminishes as the targets approach 100% proficiency requirement dictated under NCLB for 2014.[10]

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle

[10] In six of the states studied (Massachusetts, Minnesota, Rhode Island, Vermont and New Hampshire, as well as Wisconsin), an index is used that 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.

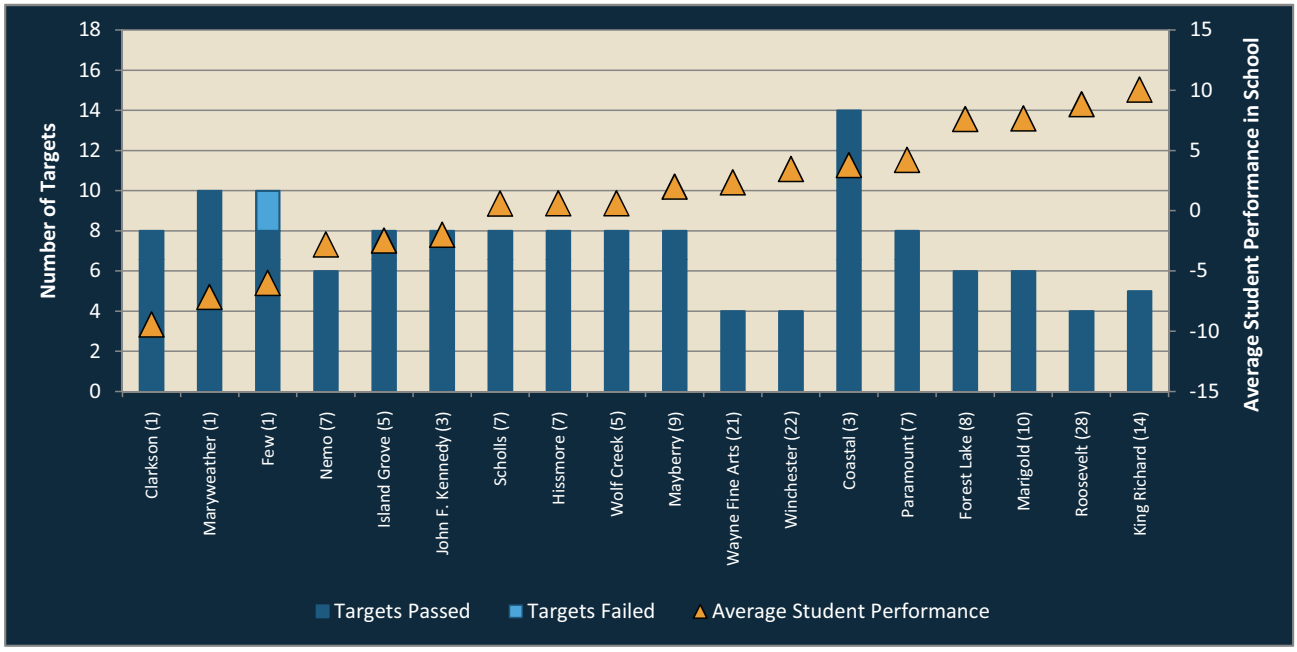

Figure 3. AYP performance of the elementary school sample under the Wisconsin 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Wisconsin's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Few Elementary, for example, met eight of its ten targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states–out of 28–in which that school would have AYP in the study.

schools. In addition, federal law requires 95% of each school's students— and 95% of students in each school's subgroups—to participate in testing.

To reiterate, then, AYP decisions in the current study are modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its annual AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Wisconsin's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Wisconsin's 2008 AYP rules. **Seventeen elementary schools made AYP, while only one (Few Elementary) failed to make it**. The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Wisconsin AYP rules. **Out of 18 middle schools in our sample, 7 made AYP**—one low-performance school (Pogesto), which has relatively few qualifying subgroups and six higher performing schools (Hoyt, Zeus, Ocean View, Walter Jones, Artemus, and Chaucer).

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the darker portions of the bars show the actual proficiency rates at each school, and the lighter

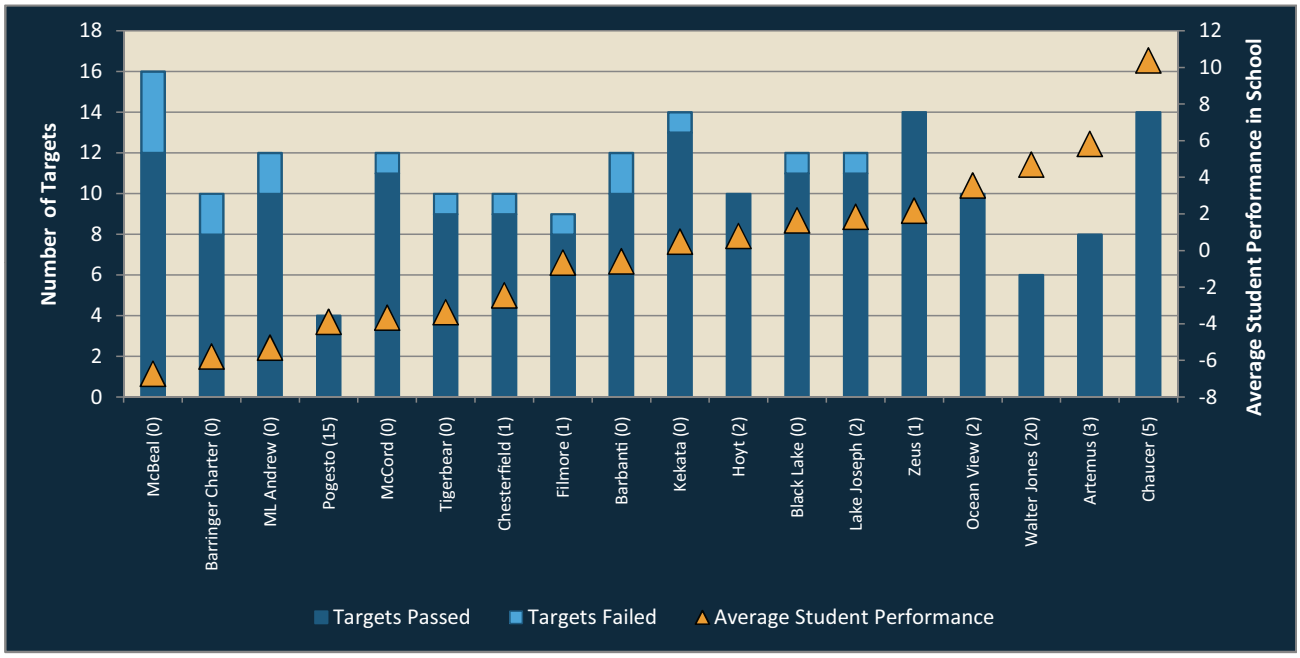

Figure 4. AYP performance of the middle school sample under the Wisconsin 2008 AYP rules

Note: This figure indicates how each middle school within the sample would have fared under Wisconsin's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup does not make AYP, so any light blue means that the school failed. Black Lake, for example, met 11 of its 12 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states, out of 28, in which that school would have made AYP.

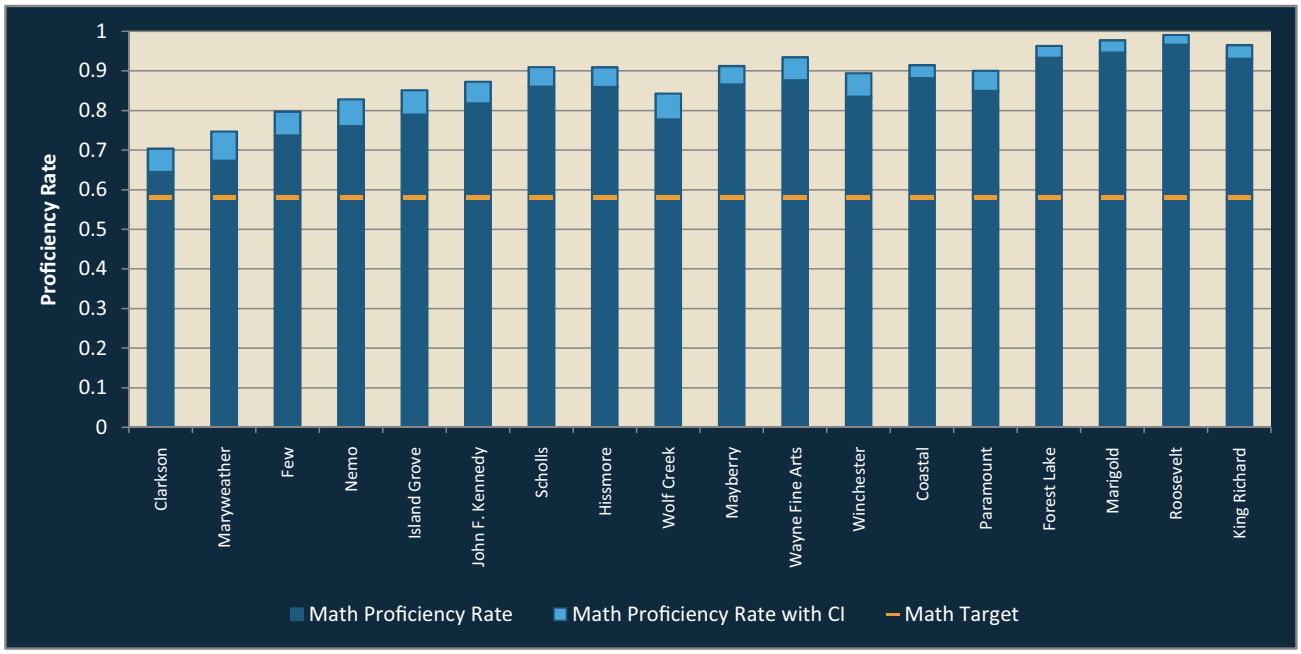

Figure 5. Impact of the confidence interval on elementary school math proficiency rates under the Wisconsin 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

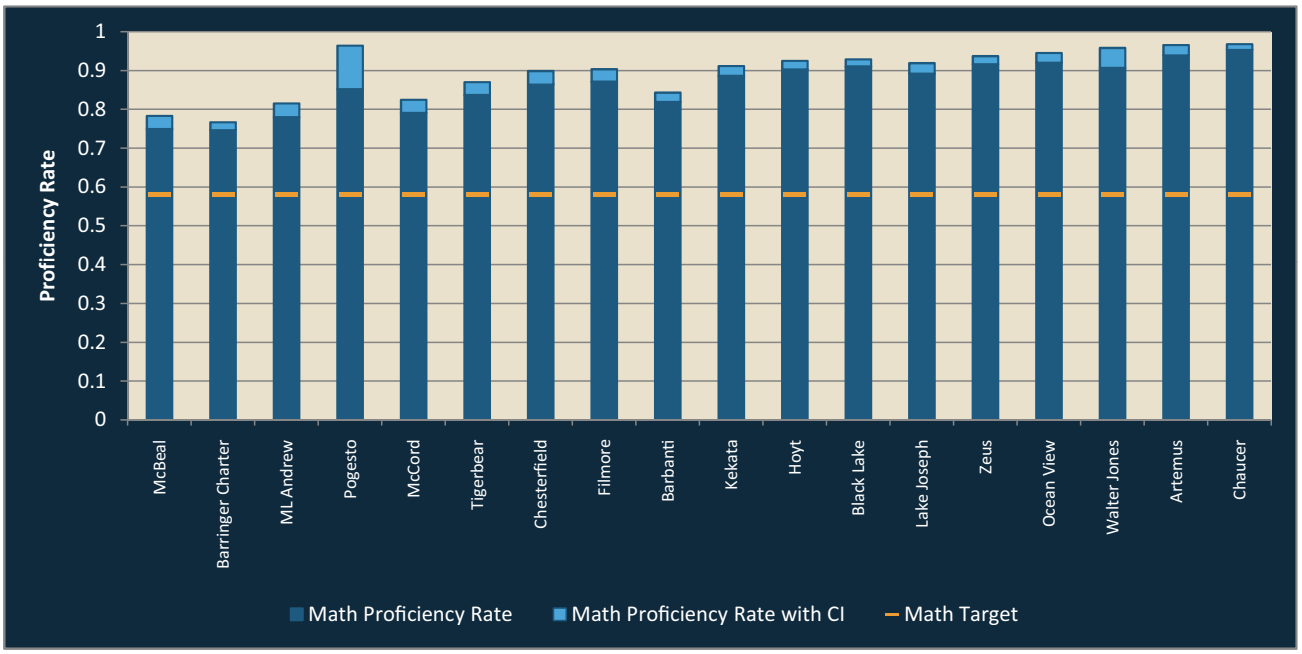

Figure 6. Impact of the confidence interval on middle school math proficiency rates under the Wisconsin 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample middle schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

portions of the bars show the degree to which these proficiency rates were increased by applying the confidence interval. The orange lines show the annual target needed to meet AYP. These figures show that **none of the sample elementary or middle schools was assisted by the confidence intervals, because the math targets in Wisconsin are so low, relative to the schools' overall performance**. The picture is much the same for reading proficiency rates at the elementary and middle school levels (not shown). No school is assisted by the confidence interval because the reading targets are so low. **In short, applying the confidence interval, even though it is a lenient one, has no effect on whether or not sample schools meet their overall reading and math targets.**[11]

Where do schools fail?

Figures 3 and 4 illustrate how many subgroup targets each sample school is held accountable, and whether or not each school made AYP. However, these figures do not indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American (AA), Asian/Pacific Islander (Asian), Hispanic/Latino (Hispanic), American Indian/Alaska Native (AI/AN), and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Wisconsin rules, and the total number of states within the study in which that school met AYP.

[11] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval may be larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary subgroup performance of sample schools under the 2008 Wisconsin AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	64.5%	79.6%	Y	Y			Y	Y	Y	Y					Y	Y					8	8	100%	Y	1
Maryweather	67.4%	79.0%	Y	Y			Y	Y	Y	Y					Y	Y			Y	Y	10	10	100%	Y	1
Few	73.7%	80.6%	Y	Y	Y	N	Y	N	Y	Y					Y	Y					10	8	80%	N	1
Nemo	76.0%	91.2%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	7
Island Grove	79.0%	89.3%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	4
JFK	81.8%	88.6%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	3
Scholls	86.0%	91.4%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Hissmore	85.9%	92.2%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	7
Wolf Creek	77.8%	90.0%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	5
Alice Mayberry	86.6%	93.4%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	9
Wayne Fine Arts	87.6%	97.7%	Y	Y															Y	Y	4	4	100%	Y	21
Winchester	83.5%	95.0%	Y	Y															Y	Y	4	4	100%	Y	22
Coastal	88.1%	90.8%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	3
Paramount	84.9%	90.7%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	7
Forest Lake	93.3%	96.6%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	8
Marigold	94.6%	96.2%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	10
Roosevelt	96.6%	99.2%	Y	Y															Y	Y	4	4	100%	Y	28
King Richard	93.1%	97.3%	Y	Y					Y										Y	Y	5	5	100%	Y	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- All schools met both their math and reading targets for their overall student populations.
- Nine of the 11 failing middle schools only missed targets for the students with disabilities subgroup.
- One middle school (Kekata) failed to make AYP only because of its LEP subgroup.
- Unlike any other state in the study, *all* of the low-income, African American, Hispanic, Asian, American Indian, and white subgroups met both their reading and math targets.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, there are very few qualifying SWD and LEP subgroups at the elementary level (keep in mind that the minimum *n* size for SWDs is rather large at 50). But when there are large enough numbers of these students to comprise subgroups at the middle school level, they tend to struggle. In fact, one of the two elementary schools and most of the middle schools in the study that have qualifying SWD subgroups fail to make AYP. Students with LEP

Table 3. Middle school subgroup performance of sample schools under the 2008 Wisconsin AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	74.9%	80.2%	Y	Y	N	N	N	N	Y	Y	Y	Y			Y	Y	Y	Y	Y	Y	16	12	75%	N	0
Barringer Charter	74.7%	88.5%	Y	Y	N	N			Y	Y	Y	Y			Y	Y					10	8	80%	N	0
ML Andrew	78.0%	89.1%	Y	Y	N	N			Y	Y	Y	Y			Y	Y			Y	Y	12	10	83%	N	0
Pogesto	85.2%	92.6%	Y	Y															Y	Y	4	4	100%	Y	15
McCord Charter	79.1%	89.5%	Y	Y	N	Y			Y	Y	Y	Y			Y	Y			Y	Y	12	11	92%	N	0
Tigerbear	83.7%	86.4%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	0
Chesterfield	86.4%	89.9%	Y	Y	Y	N			Y	Y	Y	Y							Y	Y	10	9	90%	N	1
Filmore	87.2%	92.3%	Y	Y		N			Y	Y					Y	Y			Y	Y	9	8	89%	N	1
Barbanti	82.0%	87.7%	Y	Y	N	N	Y	Y	Y	Y					Y	Y			Y	Y	12	10	83%	N	0
Kekata	88.7%	90.1%	Y	Y	Y	Y	Y	N	Y	Y	Y	Y			Y	Y			Y	Y	14	13	93%	N	0
Hoyt	90.3%	91.5%	Y	Y	Y	Y			Y	Y	Y	Y							Y	Y	10	10	100%	Y	2
Black Lake	91.1%	91.6%	Y	Y	Y	N			Y	Y	Y	Y			Y	Y			Y	Y	12	11	92%	N	0
Lake Joseph	89.2%	92.8%	Y	Y	Y	N	Y	Y	Y	Y					Y	Y			Y	Y	12	11	92%	N	2
Zeus	91.7%	91.9%	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y			Y	Y	14	14	100%	Y	1
Ocean View	92.0%	96.0%	Y	Y			Y	Y	Y	Y					Y	Y			Y	Y	10	10	100%	Y	2
Walter Jones	90.7%	94.3%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	20
Artemus	93.9%	93.9%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	3
Chaucer	95.3%	97.2%	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y			Y	Y	14	14	100%	Y	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

are also struggling somewhat to meet the state's targets; three schools with a large enough LEP population to qualify as a separate subgroup fail to meet targets for these students.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Wisconsin's NCLB accountability system is, at least somewhat, behaving like those in other states. For example, Wayne Fine Arts and Walter Jones made AYP in many of the states—21 and 20, respectively. And these schools made AYP in Wisconsin, too.

But Wisconsin is also home to a few anomalies. First, consider Clarkson and Maryweather elementary schools (see Table 2). They each failed to make AYP in 27 of the 28 states in our sample, yet made AYP in Wisconsin. In examining Table 2, we can see that Clarkson and Maryweather didn't meet the minimum numbers for the SWD subgroup, which create difficulty for so many other schools in the study. Without fewer accountable subgroups and easy proficiency standards (see Figure 2),

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Wisconsin AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	2	0	1
Students with limited English proficiency	4	0	1
Low-income students	15	0	0
African-American students	5	0	0
Asian/Pacific Islander students	0	0	0
Hispanic students	7	0	0
American Indian/Alaska Native students	0	0	0
White students	16	0	0

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Wisconsin AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	14	5	9
Students with limited English proficiency	7	1	2
Low-income students	17	0	0
African-American students	10	0	0
Asian/Pacific Islander students	1	0	0
Hispanic students	13	0	0
American Indian/Alaska Native students	1	0	0
White students	17	0	0

these schools made AYP even when schools with higher average performance failed.

Second, look at Pogesto Middle School (Figure 4). Even with its relatively low average performance it made AYP in Wisconsin, but failed to do so in 13 of 28 states. Like Clarkson and Maryweather, its AYP success in Wisconsin is most likely attributable to its relatively small number of targets (four, as shown in Table 3) along with the easy proficiency standards in Wisconsin compared to other states.

This is consistent with the patterns shown in Table 6, which compares schools that did and didn't make AYP on a number of academic and demographic dimensions. Within the sample, schools that make AYP do indeed show higher average student performance, but they also differ in the following ways: they have much smaller student populations, lower percentages of low income stu-

Table 6. Comparisons between schools that did and didn't make AYP in Wisconsin, 2008

	Elementary Schools		Middle Schools	
	Made AYP	Failed to make AYP	Made AYP	Failed to make AYP
Number of schools in sample	17	1	7	11
Average student body size	290	550	663	984
Average % low income	44	90	33	53
Average % nonwhite	38	89	26	55
Average performance†	1.65	-5.99	3.36	-2.23
Average % growth‡	114	135	105	94
Average number of targets to meet	7	10	9	12

† Student performance is measured by NWEA's MAP assessment and is expressed as an index of grade level normative performance. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance.

‡ Average growth refers to improvement from fall to spring on the NWEA MAP assessments, averaged across all students within the school. Growth is expressed as an index value relative to NWEA norms and is scaled as a percentage. Thus, 100% means that students at the school are achieving normative levels of growth for their age and grade. 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.

dents, and fewer subgroups (and thus fewer targets to meet).

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Wisconsin's AYP rules and AMOs for 2008. We found that 17 elementary schools and 7 middle schools—24 in all, from a total of 36—would have made AYP in Wisconsin. Looking across the 28 state accountability systems examined in the study, we see that Wisconsin has the greatest number of elementary schools making AYP in our sample. In addition, at 7, Wisconsin has the second highest number of middle schools making AYP in the sample (only Arizona has more). This is likely due to the easy proficiency standards in Wisconsin (the state's reading cut scores are below the 18th percentile), as well as its proficiency "index" which awards partial credit to students performing below proficient.

Because the overriding goal of the federal NCLB is to eliminate education disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Wisconsin's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students. All of the sample schools met the Wisconsin math and reading targets for their student populations as a whole. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that the size of a school's enrollment has so much influence over making AYP? Does it make sense that having fewer subgroups enhances the likelihood of making AYP? In the case of Wisconsin, is it "fair" that students receive partial credit for performing below proficient? Or that many subgroups meet their targets not because of improved performance but largely due to low cut scores? These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.

Wyoming

Executive Summary

The intent of the No Child Left Behind (NCLB) Act of 2001 is to hold schools accountable for ensuring that all their students achieve mastery in reading and math, 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 goals.

This report examines Wyoming's NCLB accountability system— particularly how its various rules, criteria, and practices result in schools either making AYP—or not making AYP. It also gauges how tough Wyoming's system is compared with other states. For this study, we selected 36 schools from various states around the nation, schools that vary by size, achievement, and diversity, among other factors, and determined whether each would make AYP under Wyoming's system as well as under the systems of 27 other states. We used school data and proficiency cut score[1] estimates from academic year 2005–2006, but applied them against Wyoming's AYP rules for academic year 2007–2008 (shortened to "2008" in this report).

Here are some key findings:

- We estimate that **16 of 18 elementary schools** and **17 of 18 middle schools** in our sample **failed to make AYP** in 2008 under Wyoming's accountability system. This high failure rate is partly explained by our sample, which intentionally includes some schools with a relatively large population of low-performing students. But it's also partially explained by Wyoming's proficiency standards which are relatively difficult, compared to other states. In addition, Wyoming's minimum subgroup size is 30, which is smaller than most other states we examined.[2] This means that more subgroups in Wyoming are held accountable for performance than in other states.
- **Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in Wyoming was exceeded in 20 other sample states (Wyoming ties 5 other states with only 2 elementary schools making AYP). In addition, Wyoming was one of 6 states with a single passing middle school in the sample (see Figure 1).**
- Many of the schools in our sample that failed to make AYP in Wyoming are meeting expected targets for their overall populations but failing because of the performance of individual subgroups, particularly students with disabilities (SWDs) and English language learners.[3]

Only two elementary schools and one middle school in our sample make AYP in 2008 under **Wyoming's** accountability system. This places Wyoming at the lower end of the state distribution in terms of the number of schools making AYP. This is likely due to the fact that Wyoming's proficiency standards are relatively difficult compared to other states. Almost all of Wyoming's cut scores are above the 40th percentile. Moreover, Wyoming's minimum subgroup size is 30, which is smaller than most other states we examined. This means that schools in Wyoming will have more accountable subgroups than would similar schools in other states. Finally, more subgroups in Wyoming's elementary schools failed to meet their reading targets than their math targets. This is probably because the proficiency standards in grades 3-6 reading are higher than those in math.

[1] A cut score is the minimum score that a student must receive on the Proficiency Assessment for Wyoming Students (PAWS) in order to be considered proficient under Wyoming's accountability system.

[2] It's important to keep in mind, however, that school size impacts minimum subgroup size (e.g., it makes sense for smaller schools to have smaller *n* sizes).

[3] It's important to note that students in subgroups not meeting the minimum *n* sizes are still included for accountability purposes in the overall student calculations; they simply are not treated as their own subgroup.

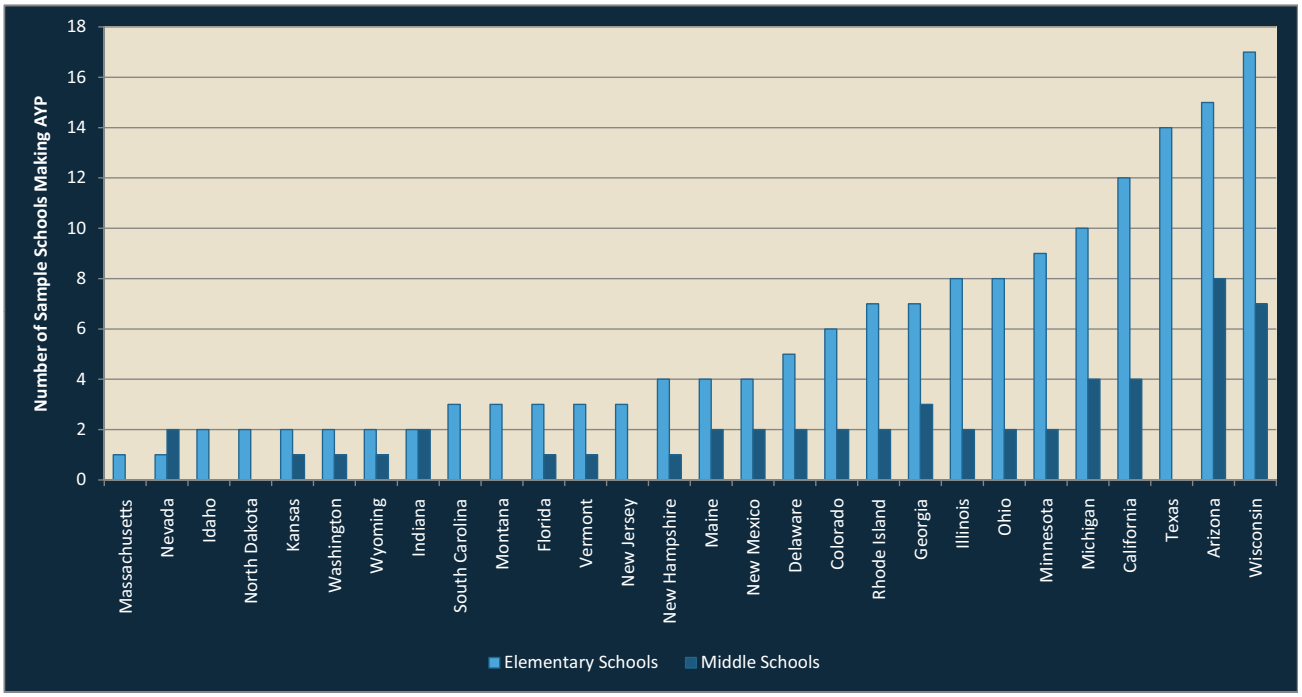

Figure 1. Number of sample schools making AYP by state

Note: Middle schools were not included for Texas and New Jersey; absence of a middle school bar in those states means "not applicable" as opposed to zero. States like Idaho and North Dakota, however, have zero passing middle schools.

- In Wyoming, as in most states, schools with fewer subgroups attain AYP more easily than schools with more subgroups, even when their average student performance is much lower. In other words, schools with greater diversity and size face greater challenges in making AYP.

- Like most other states, Wyoming applies a confidence interval (or statistical margin of error) to its measures of proficiency. **However, partly because of Wyoming's relatively low annual targets in math and reading, the confidence interval has little or no effect on AYP decisions for the sample elementary and middle schools in the state.**

- As in other states, middle schools have greater difficulty reaching AYP in Wyoming than do elementary schools, primarily because their student populations are larger and therefore have more qualifying subgroups—not because their student achievement is lower than in the elementary schools.

- A strong predictor of a school making AYP under Wyoming's system is whether it has enough English language learners to qualify as a separate subgroup. Every school with a subgroup of students with limited English proficiency (LEP)[4] failed to make AYP, in part because these students did not meet the state's targets in reading and/or math. Likewise, all schools with enough qualifying SWDs failed to meet their AYP targets in reading .[5]

Introduction

The Proficiency Illusion (Cronin et al. 2007a) linked student performance on Wyoming's tests and those of 25

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

[5] Incidentally, reading cut scores in Wyoming are higher than math cut scores in grades 3-6. In addition, SWDs are defined as those students following individualized education plans. We should also note that our subgroup findings for LEP students and SWDs may be more negative than actual findings, 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 Proficiency Assessment for Wyoming Students (PAWS), the standardized state test. 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.

other states to the Northwest Evaluation Association's (NWEA's) Measures of Academic Progress (MAP), a computerized adaptive test used in schools nationwide. This single common scale permitted cross-state comparisons of each state's reading and math proficiency standards to measure school performance under the No Child Left Behind (NCLB) Act of 2001. That study revealed profound differences in states' proficiency standards (i.e., how difficult it is to achieve proficiency on the state test), and even across grades within a single state.

Our study expands on *The Proficiency Illusion* by examining other key factors of state NCLB accountability plans and how they interact with state proficiency standards to determine whether the schools in our sample made adequate yearly progress (AYP) in 2008. Specifically, we estimated how a single set of schools, drawn from around the country, would fare under the differing rules for determining AYP in 28 states (the original 25 in *The Proficiency Illusion* plus 3 others for which we now have cut score estimates). In other words, if we could somehow move these entire schools—with their same mix of characteristics—from state to state, how would they fare in terms of making AYP? Will schools with high-performing students consistently make AYP? Will schools with low-performing students consistently fail to make AYP? If AYP determinations for schools are not consistent across states, what leads to the inconsistencies?

NCLB requires every state, as a condition of receiving Title I funding, to implement an accountability system that aims to get 100% of its students to the proficient level on the state test by academic year 2013-2014. In the intervening years, states set annual measurable objectives (AMOs). This is the percentage of students in each school, and in each subgroup within the school (such as low income[6] or African American, among others), that must reach the proficient level in order for the school to make AYP in a given year. The AMOs vary by state (as do, of course, the difficulty of the proficiency standards).

States also determine the minimum number of students that must constitute a subgroup in order for its scores to be analyzed separately (also called the minimum *n* [number of students in sample] size). The rationale is that reporting the results of very small subgroups—fewer than ten pupils, for example—could jeopardize students' confidentiality and risk presenting inaccurate results. (With such small groups, random events, like one student being out sick on test day, could skew the outcome.) Because of this flexibility, states have set widely varying *n* sizes for their subgroups, from as few as 10 youngsters to as many as 100.

Many states have also adopted confidence intervals—basically margins of statistical error—to account for potential measurement error within the state test. In some states, these margins are quite wide, which has the effect of making it easier to achieve an annual target.

All of these AYP rules vary by state, which means that a school that makes AYP in Wisconsin or Ohio, for example, might not make it under South Carolina's or Idaho's rules (U.S. Department of Education 2008.)

What We Studied

We collected students' MAP test scores from the 2005–2006 academic year from 18 elementary and 18 middle schools around the country. We also collected the NCLB subgroup designations for all students in those schools—in other words, whether they had been classified as members of a minority group, such as English language learners, among other subgroups.

The schools were not selected as a representative sample of the nation's population. Instead, we selected the schools because they exhibited a range of characteristics on measures such as academic performance, academic growth, and socioeconomic status (the latter calculated by the percentage of students receiving free or reduced-price lunches). Appendix 1 contains a complete discussion of the methodology for this project along with the characteristics of the school sample.[7]

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

[7] We gave all schools in our sample pseudonyms in this report.

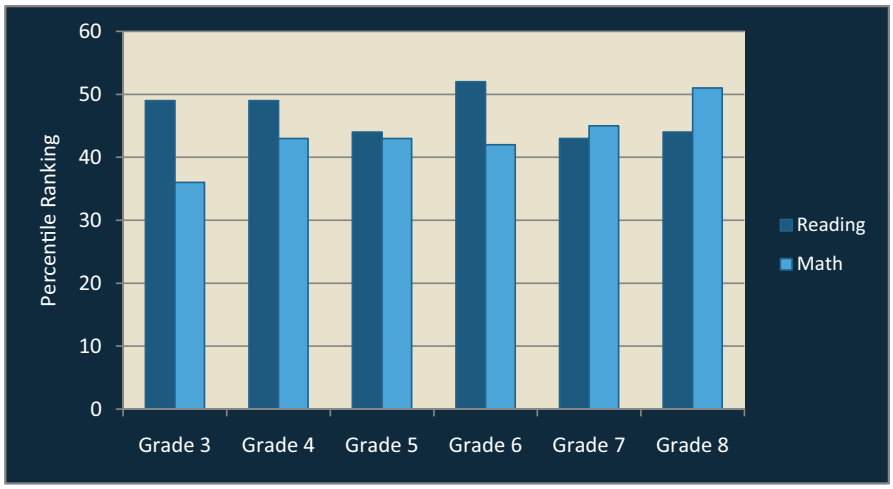

Figure 2. Wyoming reading and math cut score estimates, expressed as percentile ranks (2006)

Note: This figure illustrates the difficulty of Wyoming's cut scores (or proficiency passing scores) for its reading and math tests, as percentiles of the NWEA norm, in grades three through eight. Higher percentile ranks are more difficult to achieve. All of Wyoming's cut scores are at or below the 52nd percentile.

Proficiency cut score estimates for the Proficiency Assessment for Wyoming Students (PAWS) were estimated using the same methods as in *The Proficiency Illusion* (as shown in Figure 2), and were above average in difficulty, compared to the standards set by the other 25 states in that study. These cut scores were used to estimate whether students would have scored as proficient or better on the Wyoming test, given their performance on MAP. Student test data and subgroup designations are then used to determine how these 18 elementary and 18 middle schools would have fared under Wyoming AYP rules for 2008. In other words, the school data and our proficiency cut score estimates are from academic year 2005–2006, but we are applying them against Wyoming's 2008 AYP rules.

Table 1 shows the pertinent Wyoming AYP rules that were applied to elementary and middle schools in this study. Wyoming's minimum subgroup size is 30, which is smaller than most other states we examined.[8] As do most of the states examined in the current study, Wisconsin applies a 95% confidence interval – essentially a statistical margin of error—to their proficiency rate measurements. So, for instance, although schools are supposed to get 53.6% of their grade 3-5 students (and 53.6% of the grade 3-5 students in each subgroup) to the proficient level on the state reading test, applying the confidence interval means that the real target can actually be lower.

Note that we were unable to examine the effect of NCLB's "safe harbor" provision. This provision permits a school to make AYP even if some of its subgroups fail, as long as it reduces the number of nonproficient students within any failing subgroup by at least 10% relative to the previous year's performance. Because we had access to only a single academic year's data (2005–2006), we were not able to include this in our analysis. As a result, it is possible that some of the schools in our sample that failed to make AYP according to our estimates would have made AYP under real conditions.

Furthermore, attendance and test participation rates are beyond the scope of the study. Note that most states include attendance rates as an additional indicator in their NCLB accountability system for elementary and middle schools. In addition, federal law requires 95% of each school's students, and 95% of the students in each school's subgroup, to participate in testing.

To reiterate, then, AYP decisions in the current study are

[8] Keep in mind, however, that school size and *n* size are related (e.g., small *n* sizes make sense for small schools).

Table 1. Wyoming AYP rules for 2008

Subgroup minimum *n*	Race/ethnicity: 30	
	SWDs: 30	
	Low-income students: 30	
	LEP students: 40	
CI	**Applied to proficiency rate calculations?**	
	Yes; 95% CI used	
AMOs	**Baseline proficiency levels as of 2002 (%)**	**2008 targets (%)**
READING/LANGUAGE ARTS		
Grade 3	n/a	53.6
Grade 4	30.4	53.6
Grade 5	n/a	53.6
Grade 6	n/a	56.3
Grade 7	n/a	56.3
Grade 8	34.5	56.3
MATH		
Grade 3	n/a	49.2
Grade 4	23.8	49.2
Grade 5	n/a	49.2
Grade 6	n/a	50.2
Grade 7	n/a	50.2
Grade 8	25.3	50.2

Sources: U.S. Department of Education (2008); Council of Chief State School Officers (2008).

Abbreviations: SWDs = students with disabilities; LEP = limited English proficiency; CI = confidence interval; AMOs = annual measurable objectives; n/a=not available

modeled solely on test performance data for a single academic year. For each school, we calculated reading and math proficiency rates (along with any confidence intervals) to determine whether the overall school population and any qualifying subgroups achieved the AMOs. We deemed that a school made AYP if its overall student body and all its qualifying subgroups met or exceeded its AMOs. Again, Appendix 1 supplies further methodological detail.

How Did the Sample Schools Fare under Wyoming's AYP Rules?

Figure 3 illustrates the AYP performance of the sample elementary schools under Wyoming's 2008 AYP rules. **Only 2 elementary schools made AYP while 16 failed to make AYP.** The triangles in Figure 3 show the average academic performance of students within the school, with negative values indicating below-grade-level performance for the average student, and positive values indicating above-grade-level performance. The two passing schools (Wayne Fine Arts and Roosevelt) are in the right half of the figure, meaning that the higher performing students were found at these schools.

Yet these two schools also have relatively few qualifying subgroups—and thus the fewest targets to meet (because each subgroup has separate targets). For exam-

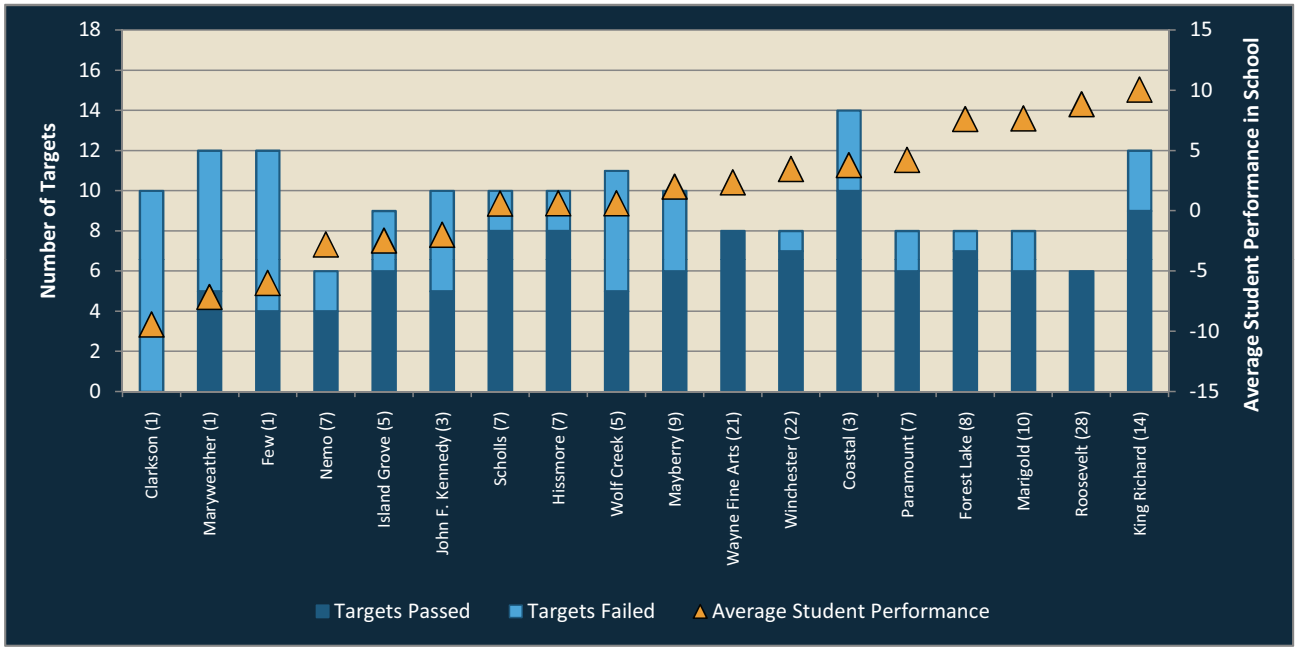

Figure 3. AYP Performance of the elementary school sample under Wyoming's 2008 AYP rules

Note: This figure indicates how each elementary school within the sample fared under Wyoming's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMOs for even a single subgroup didn't make AYP, so any light blue means that the school failed. Winchester Elementary, for example, met seven of its eight targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

ple, Wayne Fine Arts passed, but had only eight targets—two targets in reading and math for their overall student population, two more for their low-income subgroup, two more for their African American subgroup, and two for their Caucasian subgroup.

Figure 4 illustrates the AYP performance of the sample middle schools under the 2008 Wyoming AYP rules. **Out of 18 middle schools in our sample, only 1 passed**—a high-performance school (Walter Jones) which has relatively few qualifying subgroups.

Figures 5 and 6 indicate the degree to which schools' math proficiency rates are aided by the confidence interval for elementary and middle schools, respectively. On these figures, the dark blue bars show the actual proficiency rates at each school, and the light blue bars show the degree to which these proficiency rates were increased by applying the confidence interval. The orange lines show the annual measurable objective needed to meet AYP. These figures show that none of the sample elementary schools and only one of the sample middle schools (Pogesto) was assisted by the confidence interval, because the math targets in Wyoming are so low, relative to the sample schools' overall performance. Moreover, we know from Figure 4 that Pogesto still failed to meet its targets for one of its subgroups, so even though it met its overall target through use of the confidence interval, the final AYP outcome was not affected.

The effect of confidence intervals on elementary and middle school reading proficiency rates is much the same (not shown). In reading, one elementary school (Nemo) and one middle school (Filmore) were able to meet the overall target with the confidence interval, although we know from Figures 3 and 4 that these schools still failed to meet targets for subgroups. **In short, applying the**

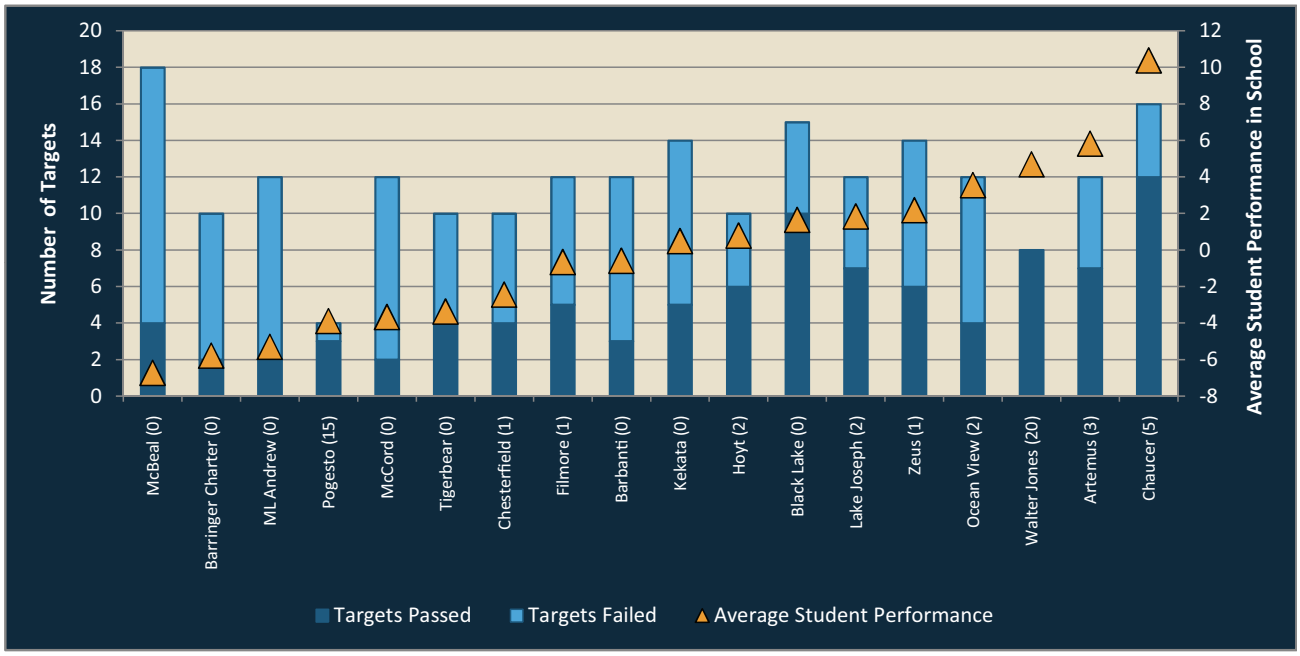

Figure 4. AYP performance of the middle school sample under Wyoming's 2008 AYP rules

Note: This figure indicates how each middle school within the sample fared under Wyoming's AYP rules (as described in Table 1). The bars show the number of targets that each school has to meet to make AYP under the state's NCLB rules, and whether they met them (dark blue) or did not meet them (light blue). The more subgroups in a school, the more targets it must meet. Under the study conditions, a school that failed to meet the AMO for even a single subgroup did not make AYP, so any light blue means that the school failed to make AYP. Hoyt, for example, met 6 of its 10 targets, but because it didn't meet them all, it didn't make AYP. Schools are ordered from lowest to highest average student performance (shown by the orange triangles), which is measured by the average MAP performance of students within the school; its scale is shown on the right side of the figure. Scores below zero (which is the grade level median) denote below-grade-level performance and scores above zero denote above-grade-level performance. One unit does not equal a grade level; however, the higher the number, the better the average performance and the lower the number, the worse the average performance. The number in parentheses after each school name indicates the number of states (out of 28) in which that school would have made AYP.

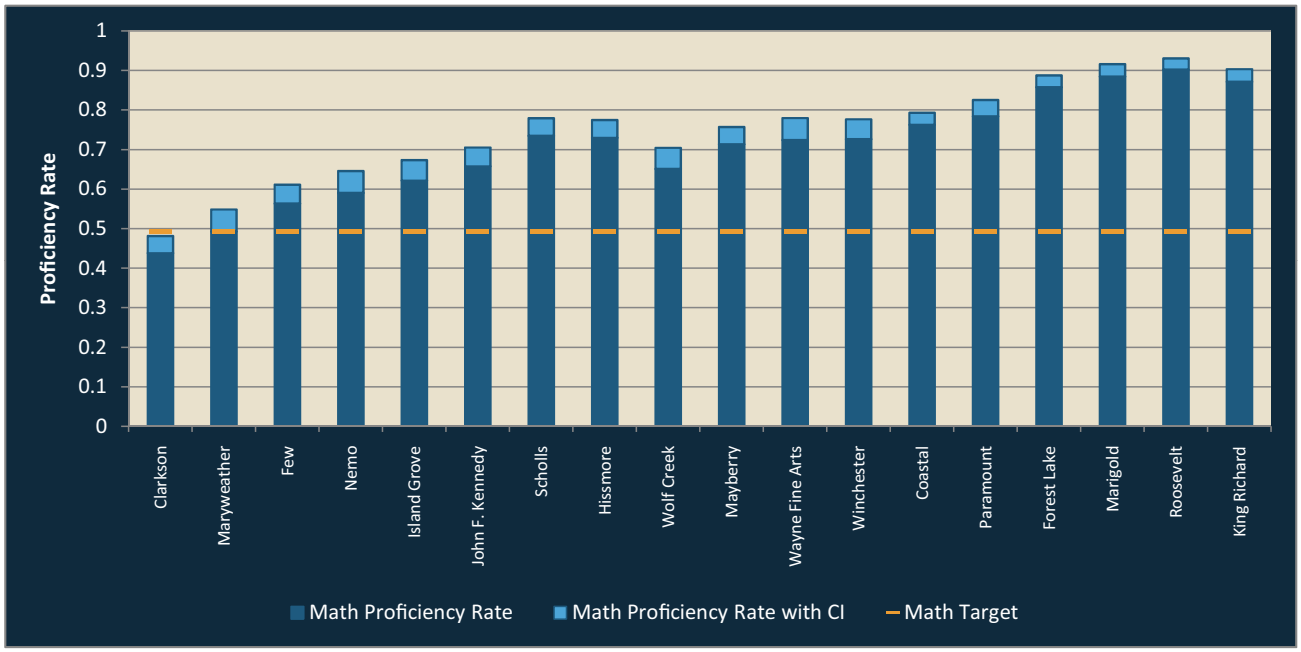

Figure 5. Impact of the confidence interval on elementary school mathematics proficiency rates under Wyoming's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that none of the sample elementary schools was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

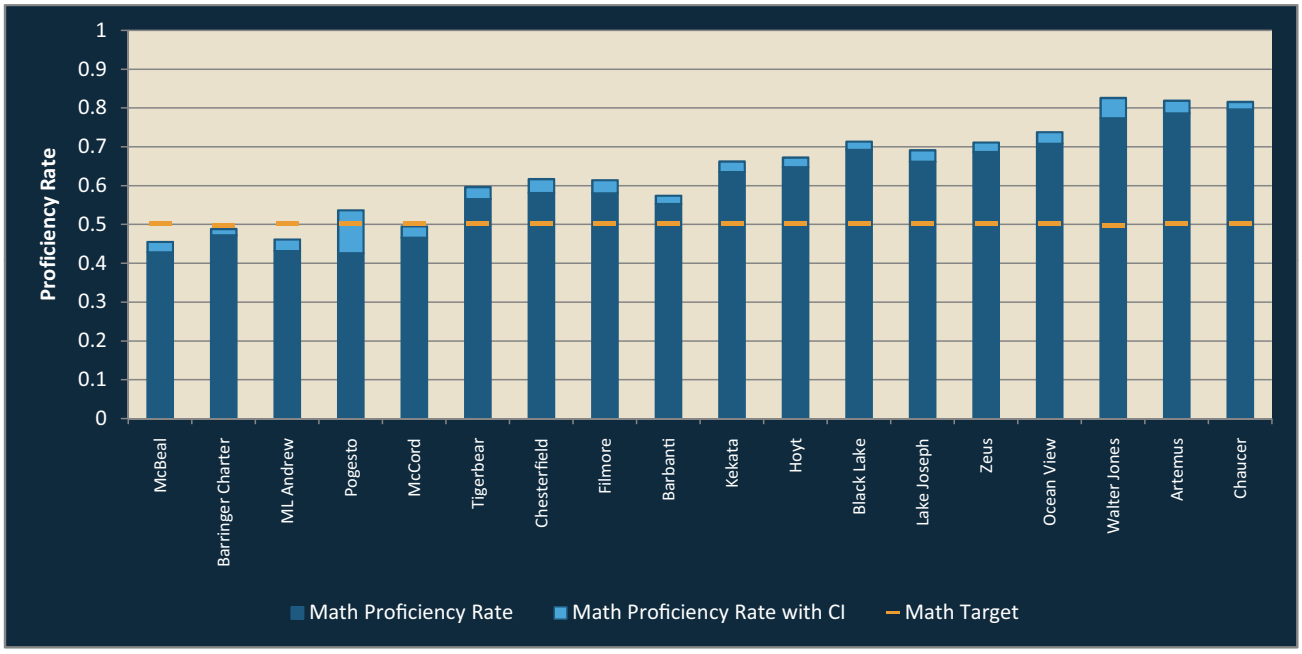

Figure 6. Impact of the confidence interval on middle school mathematics proficiency rates under Wyoming's 2008 AYP rules

Note: This figure shows the reported proficiency rate for the student population as a whole and the impact of the confidence interval on meeting annual targets. The darker portions of the bars show the actual proficiency rate achieved, while the lighter (upper) portions of the bars show the margin of error as computed by the confidence interval. The figure shows that one of the sample middle schools (Pogesto) was assisted by the confidence interval. Annual targets (the orange lines) are considered to be met by the confidence interval if they fall within the light blue portion.

confidence interval has little or no effect on AYP decisions for the sample elementary and middle schools in Wyoming.[9]

Where Do Schools Fail?

Figures 3 and 4 illustrate that schools with low or middling performance can still pass AYP when the school has fewer targets to meet because it has to fewer subgroups. These figures do not, however, indicate which subgroups failed or passed in which school. Tables 2 and 3 list information on individual subgroup performance for elementary and middle schools, respectively.

Tables 2 and 3 show which subgroups qualified for evaluation at each school (i.e., whether the number of students within that subgroup exceeded the state's minimum *n*), and whether that subgroup passed or failed. Although all schools are evaluated on the proficiency rate of their overall population, potential subgroups that are separately evaluated for AYP include SWDs, students with LEP, low-income students, and the following race/ethnic categories: African American, Asian/Pacific Islander, Hispanic/Latino, American Indian/Alaska Native, and white. Tables 2 and 3 also show whether a school met AYP under the 2008 Wyoming rules, and the total number of states within the study in which that school met AYP.

The school-by-school findings in Tables 2 and 3 show that:

- Most schools, especially at the elementary level, met their targets for their overall student population.
- Four elementary schools, however, failed to meet the reading targets for their overall school population.
- One elementary school (Clarkson) failed to meet the math targets for its overall population.
- Eight middle schools failed to meet overall proficiency targets in reading, math, or both.

[9] In the current analyses, confidence intervals were applied to both the overall school population and to all eligible subgroups in our sample schools. Thus, the ultimate impact of the confidence interval is likely larger than the impact depicted in Figures 5 and 6. However, we chose not to show how the confidence interval impacted subgroup performance because it would have added greatly to the report's length and complexity.

Table 2. Elementary school subgroup performance of sample schools under the 2008 Wyoming AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
Clarkson	43.8%	23.4%	N	N	N	N	N	N	N	N					N	N					10	0	0%	N	1
Maryweather	49.3%	35.6%	Y	N	N	N	N	N	Y	N					Y	N			Y	Y	12	5	42%	N	1
Few	56.4%	36.4%	Y	N	N	N	N	N	Y	N					Y	N			Y	N	12	4	33%	N	1
Nemo	59.1%	49.3%	Y	Y					N	N									Y	Y	6	4	67%	N	7
Island Grove	62.2%	53.1%	Y	Y				N	Y	N					Y	N			Y	Y	9	6	67%	N	4
JFK	65.8%	45.9%	Y	N	N	N			Y	N	Y	N							Y	Y	10	5	50%	N	3
Scholls	73.5%	53.9%	Y	Y	Y	N			Y	Y	Y	N							Y	Y	10	8	80%	N	7
Hissmore	73.0%	57.3%	Y	Y	N	N			Y	Y	Y	Y							Y	Y	10	8	80%	N	7
Wolf Creek	65.1%	58.9%	Y	Y	N	N		N	Y	N					N	N			Y	Y	11	5	45%	N	5
Alice Mayberry	71.4%	58.1%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	9
Wayne Fine Arts	72.4%	69.5%	Y	Y					Y	Y	Y	Y							Y	Y	8	8	100%	Y	21
Winchester	72.6%	67.3%	Y	Y	Y	N									Y	Y			Y	Y	8	7	88%	N	22
Coastal	76.3%	65.1%	Y	Y	Y	N	Y	N	Y	Y	Y	N			Y	N			Y	Y	14	10	71%	N	3
Paramount	78.4%	67.2%	Y	Y					Y	N					Y	N			Y	Y	8	6	75%	N	7
Forest Lake	85.8%	75.0%	Y	Y	Y	N			Y	Y									Y	Y	8	7	88%	N	8
Marigold	88.5%	78.3%	Y	Y	Y	N			Y	N									Y	Y	8	6	75%	N	10
Roosevelt	90.2%	84.1%	Y	Y					Y	Y									Y	Y	6	6	100%	Y	28
King Richard	87.2%	82.3%	Y	Y	Y	N	Y	N	Y	N					Y	Y			Y	Y	12	9	75%	N	14

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (Clarkson) to highest (King Richard) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

- Three of the 16 elementary schools (Hissmore, Winchester, and Forest Lake) that failed to make AYP missed only for the SWD subgroup.

Tables 4 and 5 summarize subgroup performance for elementary and middle schools, respectively. First, the performance of students with disabilities is proving challenging for schools under Wyoming's system, particularly in middle schools, where this subgroup tends to have enough students to meet the state's minimum *n* of 30. In fact, all middle schools with qualifying SWD subgroups failed to meet targets for this subgroup in both reading and math. Students with LEP are also struggling to meet the state's targets; every school with a LEP population large enough to qualify as a separate subgroup failed to meet its reading targets for these students. Low-income students also struggled; more than half of the schools with low-income subgroups failed to meet their proficiency targets for this group.

Other state reports contain a section comparing some of the characteristics of the sample schools that made AYP vs. those that did not. In Wyoming, such comparisons are less helpful, given that there were so few schools making AYP (only two elementary and one middle school).

Table 3. Middle school subgroup performance of sample schools under the 2008 Wyoming AYP rules

SCHOOL PSEUDONYM	Overall Proficiency Rate		Overall		SWDs		LEP Students		Low-income Students		AA		Asian		Hispanic		AI/AN		White		AYP Targets Required	Targets MET	% of Targets Met	School Met AYP?	Number of states in which school met AYP?
	Math	Reading	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R					
McBeal	42.8%	44.2%	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	18	4	22%	N	0
Barringer Charter	47.2%	41.8%	N	N	N	N			N	N	N	N			Y	Y					10	2	20%	N	0
ML Andrew	43.2%	45.5%	N	N	N	N			N	N	N	N			N	N			Y	Y	12	2	17%	N	0
Pogesto	42.6%	44.4%	Y	N															Y	Y	4	3	75%	N	15
McCord Charter	46.6%	51.3%	N	N	N	N			N	N	N	N			N	N			Y	Y	12	2	17%	N	0
Tigerbear	56.5%	44.9%	Y	N	N	N			Y	N	N	N							Y	Y	10	4	40%	N	0
Chesterfield	58.1%	46.8%	Y	N	N	N			Y	N	N	N							Y	Y	10	4	40%	N	1
Filmore	58.0%	55.0%	Y	Y	N	N	N	N	Y	N					N	N			Y	Y	12	5	42%	N	1
Barbanti	55.2%	53.9%	Y	N	N	N	N	N	N	N					N	N			Y	Y	12	3	25%	N	0
Kekata	63.5%	55.9%	Y	Y	N	N	N	N	Y	N	N	N			N	N			Y	Y	14	5	36%	N	0
Hoyt	64.7%	58.1%	Y	Y	N	N			Y	N	Y	N							Y	Y	10	6	60%	N	2
Black Lake	69.2%	57.5%	Y	Y	N	N	Y		Y	N	Y	N	Y	Y	Y	N			Y	Y	15	10	67%	N	0
Lake Joseph	66.2%	61.1%	Y	Y	N	N	N	N	Y	Y					Y	N			Y	Y	12	7	58%	N	2
Zeus	68.6%	61.1%	Y	Y	N	N	N	N	Y	N	Y	N			N	N			Y	Y	14	6	43%	N	1
Ocean View	70.8%	72.0%	Y	Y	N	N	N	N	N	N					N	N			Y	Y	12	4	33%	N	2
Walter Jones	77.3%	70.9%	Y	Y					Y	Y					Y	Y			Y	Y	8	8	100%	Y	20
Artemus	78.6%	69.9%	Y	Y	N	N			Y	N			Y	Y	N	N			Y	Y	12	7	58%	N	3
Chaucer	79.6%	79.4%	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	16	12	75%	N	5

Abbreviations: M = math; R = reading; N = no; Y = yes; SWDs = students with disabilities; AA = African American; Asian/Pacific Islander = Asian; Hispanic/Latino = Hispanic; American Indian/Alaska Native = AI/AN.

Note: Schools are ordered from lowest (McBeal) to highest (Chaucer) average student performance as measured by combined and weighted math and reading performance on the MAP assessment (not shown in table). A blank space underneath a subgroup means that subgroup contained fewer than the minimum number of students required for evaluation, so it wasn't counted. A "Y" in blue means that the group met the AMOs and an "N" in peach means that the group did not meet the AMOs. The two rightmost columns show (1) whether that school met AYP (i.e., it met the targets for its overall population and all required subgroups); and (2) the total number of states in the study for which that school met AYP.

In general, schools not making AYP had higher numbers of accountable subgroups than did schools making AYP, but other striking differences were not apparent.

Characteristics of Schools that Did and Didn't Make AYP

A close look at Figures 3 and 4 indicates that Wyoming's NCLB accountability system is, in many respects, behaving like those in other states. For example, among the elementary schools in our sample, Roosevelt and Wayne Fine Arts made AYP in the greatest number of states—28 and 21, respectively. And these schools made AYP in Wyoming, too. Likewise, the elementary and middle schools that fail to make AYP in the greatest number of states also fail AYP in Wyoming.

But Wyoming is home to at least one anomaly. Consider Winchester Elementary (see Figure 3). It made AYP in 22 of the 28 states in our sample, yet failed to make AYP in Wyoming. Examining Table 2, one can see that Winchester meets the minimum number (30) for the SWD subgroup, and does not meet its reading target, probably due to harder than average proficiency cut scores.

Table 4. Summary of subgroup performance of sample elementary schools under the 2008 Wyoming AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	13	7	13
Students with limited English proficiency	7	3	7
Low-income students	17	2	11
African-American students	6	0	4
Asian/Pacific Islander students	0	0	0
Hispanic students	9	2	7
American Indian/Alaska Native students	0	0	0
White students	17	0	1

Table 5. Summary of subgroup performance of sample middle schools under the 2008 Wyoming AYP rules

SUBGROUP	Number of schools with qualifying subgroups	Number of schools where subgroup failed to meet math target	Number of schools where subgroup failed to meet reading target
Students with disabilities	16	16	16
Students with limited English proficiency	9	8	8
Low-income students	17	6	14
African-American students	11	7	10
Asian/Pacific Islander students	4	0	0
Hispanic students	14	9	11
American Indian/Alaska Native students	1	1	1
White students	17	0	0

Concluding Observations

This study examined the test performance data of students from 18 elementary and 18 middle schools across the country to see how these schools would fare under Wyoming's AYP rules and AMOs for 2008. We found that only 2 elementary schools and 1 middle school—3 in all from a total of 36—would have made AYP in Wyoming. Looking across the 28 state accountability systems examined in the study, we find that the number of elementary schools making AYP in Wyoming was exceeded in 20 other sample states (Wyoming ties 5 other states with only 2 elementary schools making AYP). This is partly due to Wyoming's proficiency standards which are relatively difficult compared to other states, and Wyoming's comparatively small minimum subgroup size, meaning that more subgroups in Wyoming are likely held accountable for performance than in other states.

Because the overriding goal of the federal NCLB is to eliminate education disparities within and across states, it's important to consider whether states' annual decisions about the progress of individual schools are consistent with this aim. In some respects, Wyoming's NCLB accountability system is working exactly as Congress intended: identifying as "needing attention" schools with relatively high test score averages that mask low performance for particular groups of students, such as low-income students. Many of the sample schools met the Wyoming math and reading targets for their student populations as a whole. In the pre-NCLB era, such schools might have been considered to be effective or at least not in need of improvement, even though sizable numbers of their pupils weren't meeting state standards. Disaggregating data by race, income, and so on has made those students visible. That is surely a positive step.

Yet NCLB's design flaws are also readily apparent. Does it make sense that having fewer subgroups enhances the likelihood of making AYP? Even if actual participation guidelines for English language learners and SWDs are more generous under the current state assessment system,[10] doesn't the disproportionate failure of these students to meet Wyoming's targets indicate that a new approach is needed for holding schools accountable for the performance of these students? Yes, schools should redouble their efforts to boost achievement for ELL students and students with disabilities, as for other students, but when almost no school is able to meet the goal, perhaps that indicates that the goal is unrealistic. These will be critical considerations for Congress as it takes up NCLB re-authorization in the future.

Limitations

Although the purpose of our study was to explore how various elements of accountability systems in different states jointly affect a school's AYP status, the study will not precisely replicate the AYP outcome for every single school for several reasons. Because we projected students' state test performance from their MAP scores, and because MAP assessments—unlike state tests—are not required of all students within a school, it's possible that sampling or measurement error (or both) affected school AYP outcomes within our model. Nevertheless, for all but two of the sampled schools, our projections matched NCLB-reported proficiency ratings (in each respective state) to within 5 percentage points.

An additional limitation of the study was that it was not possible to consider NCLB's safe harbor provisions, which might have allowed some schools to make AYP even though they failed to meet their state's required AMOs. A few schools would have also passed under the new growth-model pilots currently under way in a handful of states, such as Ohio and Arizona. Others identified as making AYP in our study might actually have failed to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of some subgroup within their overall student population. At the end of the day, then, it's important to keep in mind that the number of schools that did or did not make AYP in our study do not by themselves measure the effectiveness of the entire state accountability system, of which there are many parts.

Despite these limitations, we believe that the study illuminates the inconsistency of proficiency standards

[10] See footnote 5.

and some of the rules across states. It's also useful for illustrating the challenges that states face as the requirements for AYP continue to ratchet up. The national report contains additional discussion of the study methodology and its limitations.