THE PROFICIENCY ILLUSION

John Cronin, Michael Dahlin,
Deborah Adkins, and G. Gage Kingsbury

With a foreword by
Chester E. Finn, Jr., and Michael J. Petrilli

OCTOBER 2007

THOMAS B. FORDHAM INSTITUTE

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

Table of Contents

Foreword

By Chester E. Finn, Jr., and Michael J. Petrilli

No Child Left Behind made many promises, one of the most important of them being a pledge to Mr. and Mrs. Smith that they would get an annual snapshot of how their little Susie is doing in school. Mr. and Mrs. Taxpayer would get an honest appraisal of how their local schools and school system are faring. Ms. Brown, Susie's teacher, would get helpful feedback from her pupils' annual testing data. And the children themselves would benefit, too. As President Bush explained last year during a school visit, "One of the things that I think is most important about the No Child Left Behind Act is that when you measure, particularly in the early grades, it enables you to address an individual's problem today, rather than try to wait until tomorrow. My attitude is, is that measuring early enables a school to correct problems early…measuring is the gateway to success."

So far so good; these are the ideas that underpin twenty years of sensible education reform. But let's return to little Susie Smith and whether the information coming to her parents and teachers is truly reliable and trustworthy. This fourth-grader lives in suburban Detroit, and her parents get word that she has passed Michigan's state test. She's "proficient" in reading and math. Mr. and Mrs. Smith understandably take this as good news; their daughter must be "on grade level" and on track to do well in later grades of school, maybe even go to college.

Would that it were so. Unfortunately, there's a lot that Mr. and Mrs. Smith don't know. They don't know that Michigan set its "proficiency passing score"—the score a student must attain in order to pass the test—among the lowest in the land. So Susie may be "proficient" in math in the eyes of Michigan education bureaucrats but she still could have scored worse than five-sixths of the other fourth-graders in the country. Susie's parents and teachers also don't know that Michigan has set the bar particularly low for younger students, such that Susie is likely to fail the state test by the time she gets to sixth grade—and certainly when she reaches eighth grade—even if she makes regular progress every year. And they also don't know that "proficiency" on Michigan's state tests has little meaning outside the Wolverine State's borders; if Susie lived in California or Massachusetts or South Carolina, she would have missed the "proficiency" cut-off by a mile.

Mr. and Mrs. Smith know that little Susie is "proficient." What they don't know is that "proficient" doesn't mean much. This is the proficiency illusion.

Standards-based education reform is in deeper trouble than we knew, both the Washington-driven, No Child Left Behind version and the older versions that most states undertook for themselves in the years since A Nation at Risk (1983) and the Charlottesville education summit (1989). It's in trouble for multiple reasons. Foremost among these: on the whole, states do a bad job of setting (and maintaining) the standards that matter most—those that define student proficiency for purposes of NCLB and states' own results-based accountability systems.

We've known for years that there's a problem with many states' academic standards—the aspirational statements, widely available on state websites, of what students at various grade levels should know and be able to do in particular subjects. Fordham has been appraising state standards since 1997. A few states do a super job, yet our most recent comprehensive review (2006) found that "two-thirds of schoolchildren in America attend class in states with mediocre (or worse) expectations for what their students should learn." Instead of setting forth a coherent sequence of skills and content that comprise the essential learnings of a given subject—and doing so in concrete, cumulative terms that send clear signals to educators, parents and policymakers—many states settle for nebulous, content-lite standards of scant value to those who are supposed to benefit from them.

That's a serious problem, striking at the very heart of results-based educational accountability. If the desired outcomes of schooling aren't well stated, what is the likelihood that they will be produced?

Yet that problem turns out to be just the opening chapter of an alarming tale. For we also understood that, when it comes to the real traction of standards-based education reform, a state's posted academic standards aren't the most important element. What really drives behavior, determines results, and shapes how performance is reported and understood, is the passing level—also known as the "cut score"—on the state's actual tests. At day's end, most people define educational success by how many kids pass the state test and how many fail. No matter what the aspirational statements set forth as goals, the rubber meets the road when the testing program

determines that Susie (or Michelle or Caleb or Tyrone or Rosa) is or is not "proficient" as determined by their scores on state assessments.

The advent of high-stakes testing in general, and No Child Left Behind in particular, have underscored this. When NCLB asks whether a school or district is making "adequate yearly progress" in a given year, what it's really asking is whether an acceptable number of children scored at (or above) the "proficient" level as specified on the state's tests—and how many failed to do so.

What We Asked

In the present study, we set out to determine whether states' "cut scores" on their tests are high, low, or in between. Whether they've been rising or falling (i.e., whether it's been getting harder or easier to pass the state test). And whether they're internally consistent as between, say, reading and math, or fourth and eighth grade?

One cannot answer such questions by examining academic standards alone. A state may have awesome standards even as its test is easy to pass. It could have dreadful standards, yet expect plenty of its test-takers. It might have standards that are carefully aligned from one grade to the next, yet be erratic in setting its cut scores.

To examine states' cut scores carefully, you need a yardstick external to the state itself, something solid and reliable that state-specific results and trends can be compared with. The most commonly used measuring stick is the National Assessment of Educational Progress (NAEP), yet, for reasons spelled out in the pages to follow, NAEP is a less-than-perfect benchmarking tool.

However, the Northwest Evaluation Association has a long-lived, rock-steady scale and a "Measures of Academic Progress," a computerized assessment used for diagnostic and accountability purposes by schools and school systems in many states. Not all states, to be sure, but it turns out that in a majority of them (26, to be precise), enough kids participate in MAP and the state assessment to allow for useful comparisons to be made and analyses performed.

The NWEA experts accepted this challenge and this report represents their careful work, especially that of John Cronin, Michael Dahlin, Deborah Adkins, and Gage Kingsbury. The three key questions they sought to answer are straightforward and crucial:

- How hard is it to pass each state's tests?
- Has it been getting easier or harder since enactment of NCLB?
- Are a state's cut scores consistent from grade to grade? That is, is it as hard (or easy) for a 10-year-old to pass the state's fourth-grade tests as for a 14-year-old to pass the same state's eighth-grade tests?

What We Learned

The findings of this inquiry are sobering, indeed alarming. We see, with more precision than previous studies, that "proficiency" varies wildly from state to state, with "passing scores" ranging from the 6th percentile to the 77th. We show that, over the past few years, twice as many states have seen their tests become easier in at least two grades as have seen their tests become more difficult. (Though we note, with some relief, that most state tests have maintained their level of difficulty—such as it is—over this period.) And we learn that only a handful of states peg proficiency expectations consistently across the grades, with the vast majority setting thousands of little Susies up to fail by middle school by aiming precipitously low in elementary school.

What does this mean for educational policy and practice? What does it mean for standards-based reform in general and NCLB in particular? It means big trouble—and those who care about strengthening U.S. k-12 education should be furious. There's all this testing—too much, surely—yet the testing enterprise is unbelievably slipshod. It's not just that results vary, but that they vary almost randomly, erratically, from place to place and grade to grade and year to year in ways that have little or nothing to do with true differences in pupil achievement. America is awash in achievement "data," yet the truth about our educational performance is far from transparent and trustworthy. It may be smoke and mirrors. Gains (and slippages) may be illusory. Comparisons may be misleading. Apparent problems may be nonexistent or, at least, misstated. The testing infrastructure on which so many school reform efforts rest, and in which so much confidence has been vested, is unreliable—at best. We believe in results-based, test-measured, standards-aligned accountability systems. They're the core of NCLB, not to mention earlier (and concurrent)

systems devised by individual states. But it turns out that there's far less to trust here than we, and you, and lawmakers have assumed. Indeed, the policy implications are sobering. First, we see that Congress erred big-time when NCLB assigned each state to set its own standards and devise and score its own tests; no matter what one thinks of America's history of state primacy in k-12 education, this study underscores the folly of a big modern nation, worried about its global competitiveness, nodding with approval as Wisconsin sets its eighth-grade reading passing level at the 14th percentile while South Carolina sets its at the 71st percentile. A youngster moving from middle school in Boulder to high school in Charleston would be grievously unprepared for what lies ahead. So would a child moving from third grade in Detroit to fourth grade in Albuquerque.

Moreover, many states are internally inconsistent, with more demanding expectations in math than in reading and with higher bars in seventh and eighth grade than in third and fourth (though occasionally it goes the other way), differences that are far greater than could be explained by conscious curricular decisions and children's levels of intellectual development. This means that millions of parents are being told that their eight- and nine-year-olds are doing fine in relation to state standards, only to discover later that (assuming normal academic progress) they are nowhere near being prepared to succeed at the end of middle school. It means that too little is being expected of millions of younger kids and/or that states may erroneously think their middle schools are underperforming. And it means that Americans may wrongly think their children are doing better in reading than in math—when in fact less is expected in the former subject.

While NCLB does not seem to be fueling a broad "race to the bottom" in the sense of many states lowering their cut scores in order to be able to claim that more youngsters are proficient, this study reveals that, in several instances, gains on state tests are not being matched by gains on the Northwest Evaluation Association test, raising questions about whether the state tests are becoming easier for students to pass. The report's authors describe this as a "walk to the middle," as states with the highest standards were the ones whose estimated passing scores dropped the most.

NCLB aside, what is the meaning of a "standard" if it changes from year to year? What is the meaning of measurable academic gains—and "adequate yearly progress"—if the yardstick is elastic?

Standards-based reform hinges on the assumption that one can trust the standards, that they are stable anchors to which the educational accountability vessel is moored. If the anchor doesn't hold firm, the vessel moves—and if the anchor really slips, the vessel can crash against the rocks or be lost at sea.

That, we now see clearly, is the dire plight of standards-based reform in the United States today.

Looking Ahead

What to do? First, it's crazy not to have some form of national standards for educational achievement—stable, reliable, cumulative, and comparable. That doesn't mean Uncle Sam should set them, but if Uncle Sam is going to push successfully for standards-based reform he cannot avoid the responsibility of ensuring that they get set. NCLB edition 1.0 didn't do that and, so far as one can read the policy tea-leaves and bill drafts today, version 2.0 won't either. If the feds won't act, the states should, by coming together to agree to common, rational, workable standards (as most states have been doing with regard to high-school graduation rates.)

Yet even if national or inter-state standards are not in the cards in the foreseeable future, state standards clearly need an immediate and dramatic overhaul. In our view, the place to start isn't third grade; it's the end of high school. Education standards in the U.S. should be tethered to real-world expectations for the skills and knowledge that 18-year-olds need to possess in order to succeed in a modern economy and democratic polity. High-school graduation should be attached to reasonable attainment of those standards; the existing American Diploma Project is a good example of what they might look like, at least in English and math.

Then everything else should be "backward mapped" so that standards in the various grades proceed cumulatively from kindergarten to graduation and it becomes possible to know whether a child is or is not "on course" to meet the 12th-grade exit expectations. Satisfactory progress means staying on that trajectory from year to year. If Susie is behind, then she's got extra learning to do and extra efforts should be made to see that she gets the help she needs.

The "discussion draft" reauthorization proposal recently advanced by Chairman George Miller and Ranking Member Buck McKeon of the House Education and Labor committee shows faint hints of such a strategy, with financial incentives for states that adopt "world-class" standards that imply

readiness for work or college. Yet they undermine this objective by slavishly clinging to the "100 percent proficient by 2014" mandate. Policy groups from left, right, and center, including the estimable and hawkish Education Trust, now agree: this lofty aspirational objective is doing more harm than good. It has worsened the proficiency illusion. If Congress wants states like Michigan to aim higher, so that Mr. and Mrs. Smith know how Susie is really performing, the best thing it can do is to remove this provision from the law. With this perverse incentive out of the way, Michigan just might summon the intestinal fortitude to aim higher—and shoot straighter.

This, we submit, is how to begin thinking afresh about standards-based reform in general and NCLB in particular. For this enterprise not to collapse, we need standards and tests that are suitably demanding as well as stable, cumulative (all the way through high school), trustworthy, and comparable. American k-12 education is a long way from that point today.

Many people played critical roles in the development of this report. First, we thank the Joyce Foundation, and our sister organization, the Thomas B. Fordham Foundation, for the financial resources to make this ambitious project possible. Next, we appreciate the members of our advisory panel, who provided keen suggestions on our methodology, expert feedback on our drafts, and sundry recommendations that no doubt made this study a stronger product. (Of course, we accept any faults of the research or presentation as our own.) They include Andrew Porter (now at the University of Pennsylvania); Stanford's Margaret Raymond; Martin West (at Brown); and the Education Trust's Ross Wiener.

This project required immense effort to document and validate the assessment information from the twenty-six states included in this study. We thank Nicole Breedlove who contributed several months of her time and talent to this work. The final report contains over one thousand numbers, each of which had to be cross-checked and validated against their original computations, which were scattered through scores of spreadsheets and SPSS printouts. Jane Kauth contributed quality assurance expertise and experience to this task, and we greatly appreciate her contribution to the integrity of the report.

Fordham Institute staff and interns spent countless weeks proofing and editing the report; we thank Heather Cope, Martin Davis, Christina Hentges, Jeffrey Howard, Liam Julian, Amanda Klein, and Coby Loup for their efforts. Anne Himmelfarb expertly copy-edited the main part of this report; Bill Buttaggi is responsible for its clean, readable design. We appreciate all of their efforts.

Executive Summary

At the heart of the No Child Left Behind Act (NCLB) is the call for all students to be "proficient" in reading and mathematics by 2014. Yet the law expects each state to define proficiency as it sees fit and design its own tests. This study investigated three research questions related to this policy:

1. How consistent are various states' expectations for proficiency in reading and mathematics? In other words, is it harder to pass some states' tests than others?

2. Is there evidence that states' expectations for proficiency have changed since NCLB's enactment? If so, have they become more or less difficult to meet? In other words, is it getting easier or harder to pass state tests?

3. How closely are proficiency standards calibrated across grades? Are the standards for earlier grades equivalent in difficulty to those for later grades (taking into account obvious grade-linked differences in subject content and children's development)? In other words, is a state's bar for achievement set straight, sloping, or uneven?

This study used data from schools whose pupils participated both in state testing and in assessment by the Northwest Evaluation Association (NWEA) to estimate proficiency cut scores (the level students need to reach in order to pass the test for NCLB purposes) for assessments in twenty-six states. Here are the results:

• State tests vary greatly in their difficulty. Our study's estimates of proficiency cut scores ranged from the 6th percentile on the NWEA scale (Colorado's grade 3 mathematics standards) to the 77th percentile (Massachusetts' 4th grade mathematic standards). Among the states studied, Colorado, Wisconsin, and Michigan generally have the lowest proficiency standards in reading, while South Carolina, California, Maine, and Massachusetts have the highest. In math, Colorado, Illinois, Michigan, and Wisconsin have the lowest standards, while South Carolina, Massachusetts, California, and New Mexico have the highest.

• Most state tests have not changed in difficulty in recent years. Still, eight states saw their reading and/or math tests become significantly easier in at least two grades, while only four states' tests became more difficult. The study estimated grade-level cut scores at two points in time in nineteen states. Half of these cut score estimates (50 percent in reading, 50 percent in mathematics) did not change by more than one standard error. Among those that did change significantly, decreases in cut score estimates (72 percent in reading, 75 percent in mathematics) were more common than increases (28 percent in reading, 25 percent in mathematics). In reading, cut score estimates declined in two or more grades in seven states (Arizona, California, Colorado, Illinois, Maryland, Montana, and South Carolina), while cut score estimates rose in New Hampshire, New Jersey, and Texas. In mathematics, cut score estimates declined in at least two grades in six states (Arizona, California, Colorado, Illinois, New Mexico, and South Carolina) while rising in Minnesota, New Hampshire, and Texas. The declines were greatest for states that previously had the highest standards, such as California and South Carolina. Several factors could have explained these declines, which resulted from learning gains on the state test not being matched by learning gains on the Northwest Evaluation Association test.

• Improvements in passing rates on state tests can largely be explained by declines in the difficulty of those tests. This study found that the primary factor explaining improvement in student proficiency rates in many states is a decline in the test's estimated cut score. Half of the reported improvement in reading, and 70 percent of the reported improvement in mathematics, appear idiosyncratic to the state test. A number of factors could explain why our estimates of cut scores might decline, including "teaching to the state test," greater effort by students on state tests than on the NWEA exam, or actual changes to the state test itself. Regardless, these declines raise questions about whether the NCLB-era achievement gains reported by many states represent true growth in student learning.

• Mathematics tests are consistently more difficult to pass than reading tests. The math standard bests the reading standard in the vast majority of states studied. In seven states (Colorado, Idaho, Delaware, Washington, New Mexico, Montana, and Massachusetts), the difference between the eighth-grade reading and mathematics cut scores was greater than 10 percentile points. Such a discrepancy in expectations can yield the impression that students are performing better in reading than in math when that isn't necessarily the case.

• Eighth-grade tests are consistently and dramatically more difficult to pass than those in earlier grades (even after taking into account obvious differences in subject-matter complexity and children's academic development). Many states are setting the bar significantly lower in elementary school than in middle school, giving parents, educators, and the public the false impression that younger students are on track for future success—and perhaps setting them up for unhappy surprises in the future. This discrepancy also gives the public the impression that elementary schools are performing at much higher levels than middle schools, which may not be true. The differences between third-grade and eighth-grade cut scores in reading are *20 percentile points* or greater in South Carolina, New Jersey, and Texas, and there are similar disparities in math in New Jersey, Michigan, Minnesota, North Dakota, and Washington.

Thus, five years into implementation of the No Child Left Behind Act, there is no common understanding of what "proficiency" means. Its definition varies from state to state, from year to year, from subject to subject, and from grade level to grade level. This suggests that the goal of achieving "100 percent proficiency" has no coherent meaning, either. Indeed, we run the risk that children in many states may be nominally proficient, but still lacking the education needed to be successful on a shrinking, flattening, and highly competitive planet.

The whole rationale for standards-based reform was that it would make expectations for student learning more rigorous and uniform. Judging by the findings of this study, we are as far from that objective as ever.

Introduction

At the heart of the No Child Left Behind Act (NCLB) is the call for all American school children to become "proficient" in reading and mathematics by 2014. Yet that law expects each state to define proficiency as it sees fit and to design its own tests. This study investigated three research questions related to this policy.

1. How consistent are the various states' expectations for "proficiency" in reading and mathematics? Prior studies have found great variability, usually by comparing student performance on state assessments to student performance on the National Assessment of Educational Progress (NAEP). This was the approach of a June 2007 study by the National Center for Educational Statistics (NCES), *Mapping 2005 State Proficiency Standards Onto the NAEP Scale.* Yet the use of NAEP has limits. NAEP assesses students only at three grade levels: 4, 8, and 12. Because NAEP does not report individual- or school-level results, there are questions about the degree of motivation that children bring to the assessment (Educational Testing Service 1991; O'Neill et al. 1997). Finally, because NAEP is intended to be a national test, the content of the exam may not always align with that of state assessments. To address this concern, the current study used the Measures of Academic Progress (MAP) assessment, a computerized-adaptive test developed by the Northwest Evaluation Association (NWEA) and used in schools nationwide, to estimate proficiency cut scores for twenty-six states' assessments. (Proficiency cut scores are the levels that students need to reach in order to pass the test for NCLB purposes.) The use of the MAP assessment allowed us to estimate standards in grades 3 through 8. Because the MAP test reports individual results to parents and is used by school systems for both instructional and accountability purposes, students and teachers have incentives for students to perform well. Finally, the test is aligned to individual states' curriculum standards, which should improve the accuracy of cut score estimates.

2. Is there evidence that states' expectations for "proficiency" have changed over time, in particular during the years immediately following enactment of NCLB? If so, have they become more or less difficult to meet? Is it getting easier or harder to pass state tests? To determine whether states have made progress in helping more of their pupils achieve proficiency in reading or math, it is important to know whether each state's definition of proficiency has remained constant. NCLB allows states to revise their academic standards, adopt new tests, or reset their passing scores at any time. All of these changes provide opportunities for the proficiency standards to rise or fall as a result of conscious decisions or policy changes. Moreover, unintended drift in these standards may also occur over time.

3. How closely are a state's proficiency standards calibrated across grades? Are the standards in earlier grades equivalent in difficulty to proficiency standards in later grades (taking into account the obvious differences in subject content and children's development from grade to grade)? A calibrated proficiency standard is one that is relatively equal in difficulty across all grades. Thus, the eighth-grade standard would be no more or less difficult to achieve for eighth-graders than the fifth-grade or third-grade standards would be for fifth- or third-graders, respectively. When standards are calibrated in this way, parents and educators have some assurance that attaining the third-grade proficiency standard puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades result from differences in children's actual educational attainment and not simply from differences in the difficulty of the test. We examined the degree to which state proficiency standards live up to this ideal.

Methodology

This section offers a brief overview of the methods used to conduct this study. Appendix 1 contains a complete description of the our methodology.

Estimating proficiency cut scores requires that data from one measurement scale be translated to another scale that is trying to measure the same thing. Assume that we have decided that a proficient long jumper in sixth grade should be able to jump eight feet, and that we want to know how that proficiency would be expressed in meters. Because the relationship between the English and metric scales is known, this conversion is quite simple, so a single calculation allows us to know that the metric equivalent of 8 feet is 2.43 meters.

Unfortunately, the task of estimating proficiency cut scores is not quite as simple, for two reasons. First, because each state has its own proficiency test, we must compare each of the state test scales to all of the others to know the relative difficulty of each test; we cannot simply compare one scale to a second. Second, because it is not possible to make visual comparisons of the scales used to measure educational achievement (as it is with those that measure distance), we have to infer the relationship between the two scales.

We do this by comparing the performance of the same students on the two instruments. Extending the long-jump analogy, imagine that we were able to determine that 50 percent of sixth-grade long jumpers could jump eight feet, and we wanted to find the metric equivalent without knowing the conversion formula. One way to get an estimate would be to ask that same group of sixth-graders to jump a second time and measure their performance using a metric tape measure. We could then rank the results and use the 50th percentile score to estimate the point that is equivalent to eight feet. While the result might not be exactly 2.43 meters, it would generally be reasonably close to it, as long as the students performed the task under similar conditions.

This kind of process, called an equipercentile equating procedure, is commonly used to compare the scales employed on achievement tests, and it allowed us to estimate the cut scores for twenty-six state instruments on a single scale. This study used data collected from schools whose students participated both in state testing and in the NWEA MAP assessment, using the NWEA scale as a common ruler. For nineteen of these states, estimates of the proficiency cut scores could be made at two points in time (generally 2002-03 and 2005-06). These were used to look for changes that may have occurred during the process of implementing the No Child Left Behind Act. (The twenty-four excluded states did not have enough students in the NWEA sample to be included in this study.)

Instruments

State proficiency cut score equivalents were estimated using the MAP assessments, which are tests of reading and mathematics produced by NWEA and used by 2,570 school systems across forty-nine states. NWEA develops all its assessments from large pools of items that have been calibrated for their difficulty. These pools contain approximately fifty-two hundred items in reading and eight thousand items in mathematics. To create reading and math assessments for each state, NWEA curriculum experts evaluate the particular state's content standards and cross-reference each standard to an index of the NWEA item pool. About two thousand aligned items are selected for that state's final MAP assessment. Because the items drawn from each individual state assessment are all linked to a single common scale, results of the various state MAP assessments can be compared to one another.

Students taking MAP receive a test that is forty to fifty-five items in length. Each test contains a balanced sample of questions testing the four to eight primary standards in that state's curriculum. The assessment is designed to be adaptive, meaning that high- and low-performing students will commonly respond to items that are aligned to the state's content standards, but are offered at a level of difficulty that reflects the student's current performance rather than the student's 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.

Prior studies have found that student performance on MAP is closely correlated with student performance on state assessments in reading and mathematics (Northwest Evaluation Association, 2005a). These results show that the procedures used to align the content of MAP to state standards result in a test that measures similar content. A more detailed discussion of MAP is included in Appendix 1 under "Instruments."

Cut Score Estimation Procedure

For purposes of this study, we use the term "proficiency cut score" to refer to the score on each state's assessment that is used to report proficient performance for the purposes of the No Child Left Behind Act. Two states in this study have not always used the "proficient" level on their state test to represent proficiency for NCLB. Colorado uses the "partially proficient" level of performance on its state test for this purpose, and New Hampshire, prior to its adoption of the New England Common Assessment Program (NECAP), used the "basic" level of performance to report proficiency. Today, New Hampshire uses the "proficient" level of performance on NECAP for NCLB reporting.

To estimate the difficulty of each state's proficiency cut scores for reading and mathematics, we linked results from state tests to results from the NWEA assessment. In fifteen states, this was done by analyzing a group of schools in which almost all students had taken both the state's assessment and the NWEA test. In the other eleven states, we had direct access to student-level state assessment results. In these states, the researchers matched the state test result for each student directly to his or her MAP results to form the sample used to generate the cut score estimate. These sampling procedures identified groups of students in which nearly all participants took both MAP and their respective state assessment. A more detailed discussion of the procedures used to create the population sample is included in Appendix 1 under "Sampling."

To estimate proficiency-level cut scores, the researchers found the proportion of students within the sample who achieved at the proficient level or better on the state assessment. Following the equipercentile method, they then found the score on the NWEA scale that would produce an equivalent proportion of students. For example, if 75 percent of the students in the sample achieved proficient performance on their state assessment, then the score of the 25th percentile student in the sample (100 percent of the group minus the 75 percent of the group who achieved proficiency) would represent the minimum score on MAP associated with proficiency on the state test. The methods used in this study to estimate proficiency-level cut scores were evaluated in a preliminary study and found to predict state-test result distributions with a high level of accuracy (Cronin et al. 2007). A more detailed discussion of the methods used to estimate cut scores can be found in Appendix 1 under "Estimates."

All estimates of cut scores were made directly to the NWEA scale. To make comparisons easier for readers, scale scores were converted to percentiles for reporting purposes.

Cut score estimates were used in three types of comparisons. First, the most recent cut score estimate was used to compare the difficulty of proficiency standards across the twenty-six states in the study. For some grade levels, we were not able to estimate cut scores for all twenty-six states, generally because of insufficient sample size. Second, the most recent cut score estimate was also compared to a prior cut score estimate for nineteen states in reading and eighteen states in mathematics in an effort to determine how the difficulty of standards may have changed during the study period. (The NWEA scale is stable over time.) Third, the researchers examined differences in the difficulty of cut score estimates between grades within each state. This was done in an effort to determine whether performance expectations for the various grades were consistent.

These comparisons permitted us to answer the three major questions of the study: 1) How consistent are the various states' expectations for proficiency in reading and mathematics? 2) Is there evidence that states' expectations for proficiency have changed over time? 3) How closely are proficiency standards calibrated across grades? That is, are the standards in earlier grades equal in difficulty to proficiency standards in later grades?

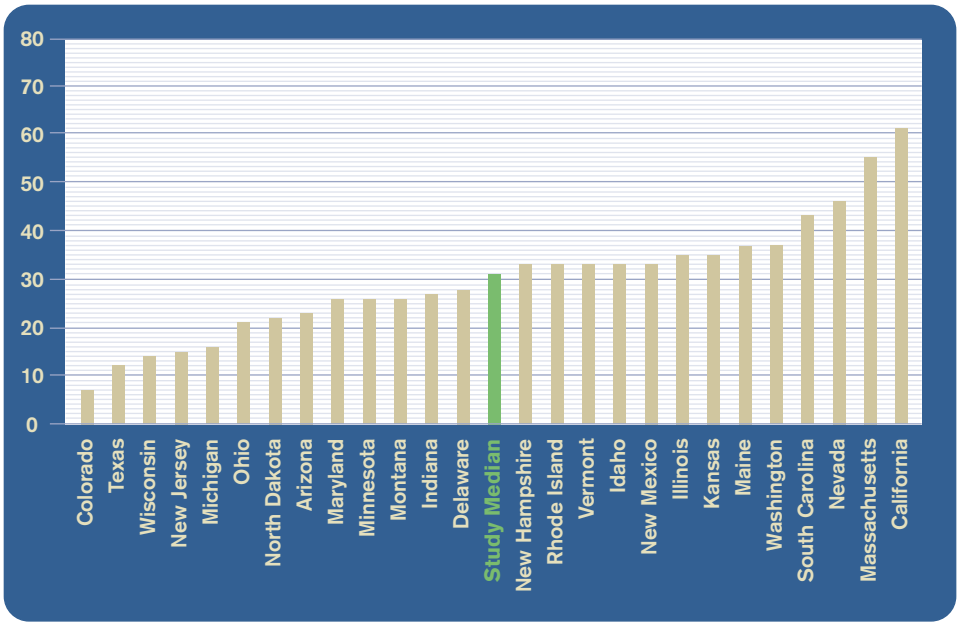
National Findings

Question 1:
How consistent are the various states' expectations for "proficiency" in reading and mathematics?

State tests vary greatly in their difficulty.

Figure 1 depicts grade 3 reading proficiency cut score estimates used for NCLB purposes in each of the twenty-six states studied. (Individual grade results for each state appear in Appendices 4 and 5.) These ranged from the 7th percentile (Colorado) to the 61st percentile (California) on the NWEA scale. In twenty-four of the twenty-six states examined, the grade 3 proficiency cut score was below the 50th MAP percentile, with nineteen of the twenty-six estimated cut scores falling in the second quintile, or the 20th to 40th percentile range.

Figure 1 – Grade 3 estimated reading proficiency cut scores for 2006 (ranked by MAP percentile)



Note: This figure ranks the grade 3 reading cut scores from easiest (Colorado) to most difficult (California) and shows the median difficulty across all states studied (in green).

Colorado currently reports the state's "partially proficient" level of academic performance on its state test as "proficient" for NCLB purposes, while using the higher "proficient" level for internal state evaluation purposes. In effect, Colorado has two standards: an easier standard for NCLB, and a harder standard for internal state use. For purposes of fairly comparing Colorado to other states, we used their NCLB-reported standard. Consequently, all subsequent references to "proficient" or "proficiency" in Colorado should be understood as referring to the NCLB-reported standard.

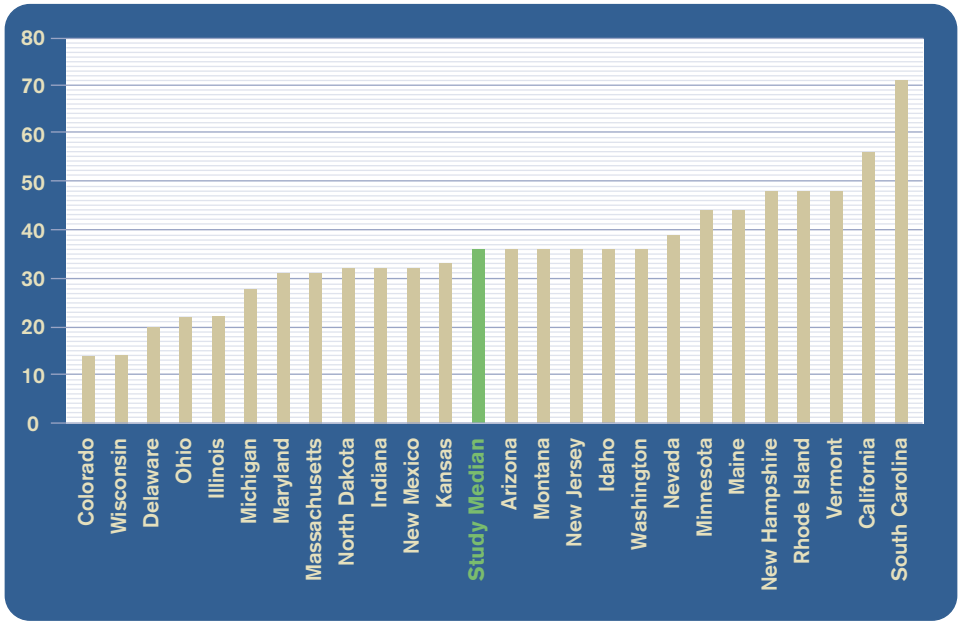
Figure 2 depicts the range of grade 8 reading proficiency cut scores for twenty-five of the states studied. Eighth-grade scores ranged from the 14th percentile (Colorado) to the 71st percentile (South Carolina) on the NWEA scale. Eighth-grade proficiency cut scores were less clustered than the third-grade scores. In twenty-three of the twenty-five states examined, the average score required for proficiency was below the 50th percentile, and sixteen of the twenty-five states' estimated cut scores fell in the second quintile.

Figure 3 depicts the range of grade 3 math proficiency cut scores in each of the twenty-five states studied (excluding Maryland, which used the NWEA MAP test only for reading). The mathematics standards show greater variability than the reading standards, ranging in difficulty from the 6th percentile (Colorado and Michigan) to the 71st percentile (South Carolina). The proficiency cut scores of twenty-two of the twenty-five states were below the 50th percentile, and thirteen fell into the second quintile.

Figure 4 depicts grade 8 math proficiency cut scores in twenty-two states. They range in difficulty from the 20th percentile (Illinois) to the 75th percentile (South Carolina). The eighth-grade standards were above the 50th percentile in ten states, and the cut score estimates for nine of the remaining twelve states were in the second quintile.

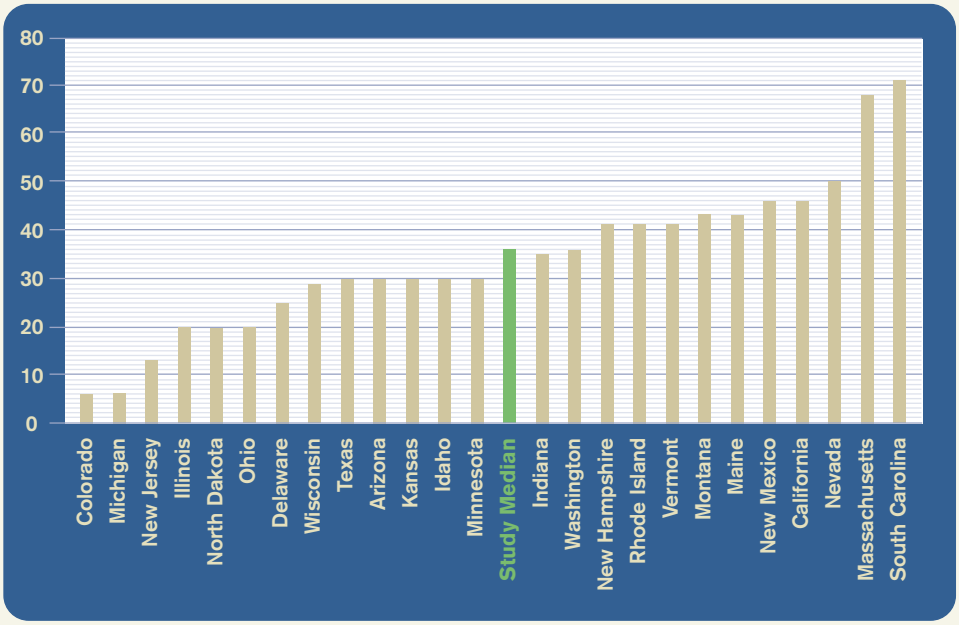
Figures 5 and 6 show the average rank of state cut scores across all grades, where the lowest rank reflects the least difficult cut score and the highest rank denotes the most difficult. In reading (Figure 5), we found that Maine, California, and South Carolina generally had the highest proficiency cut scores, while Colorado, Wisconsin, and Michigan had the lowest. In math (Figure 6), California, Massachusetts, and South Carolina had the highest proficiency cut scores, while Colorado, Illinois, and Michigan had the lowest, on average.

Figure 2 – Grade 8 estimated reading proficiency cut scores for 2006 (ranked by MAP percentile)



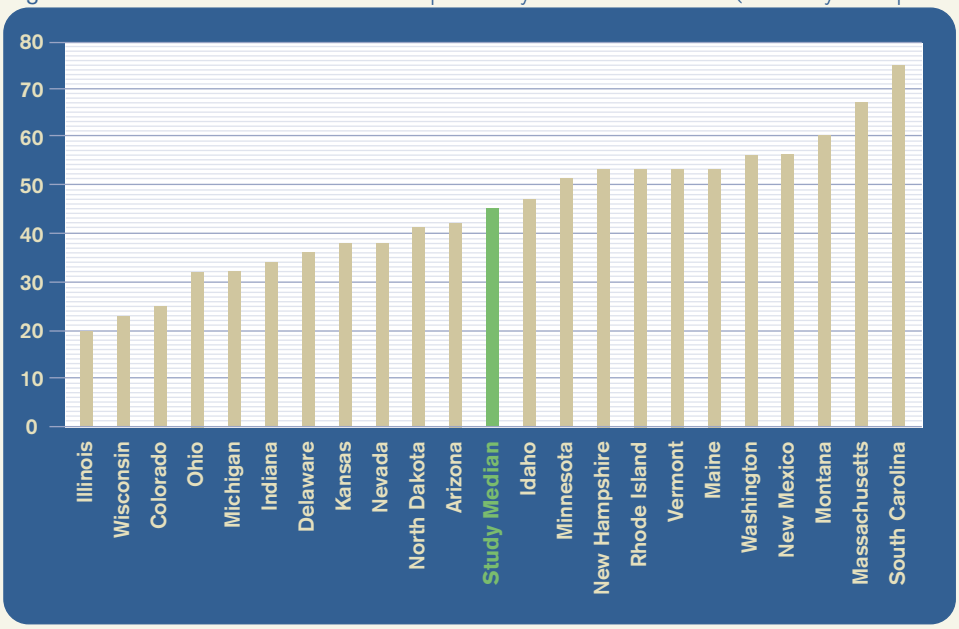
Note: This figure ranks the grade 8 reading cut scores from easiest (Colorado) to most difficult (South Carolina) and shows the median difficulty across all states studied (in green).

Figure 3 – Grade 3 estimated mathematics proficiency cut scores for 2006 (ranked by MAP percentile)



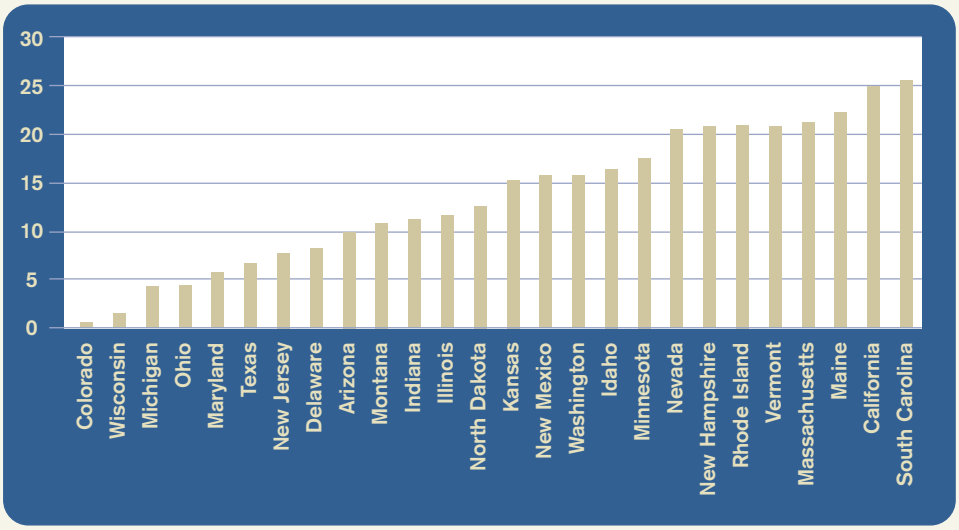
Note: This figure ranks the grade 3 math cut scores from easiest (Colorado) to most difficult (South Carolina) and shows the median difficulty across all states studied (in green).

Figure 4 – Grade 8 estimated mathematics proficiency cut scores for 2006 (ranked by MAP percentile)



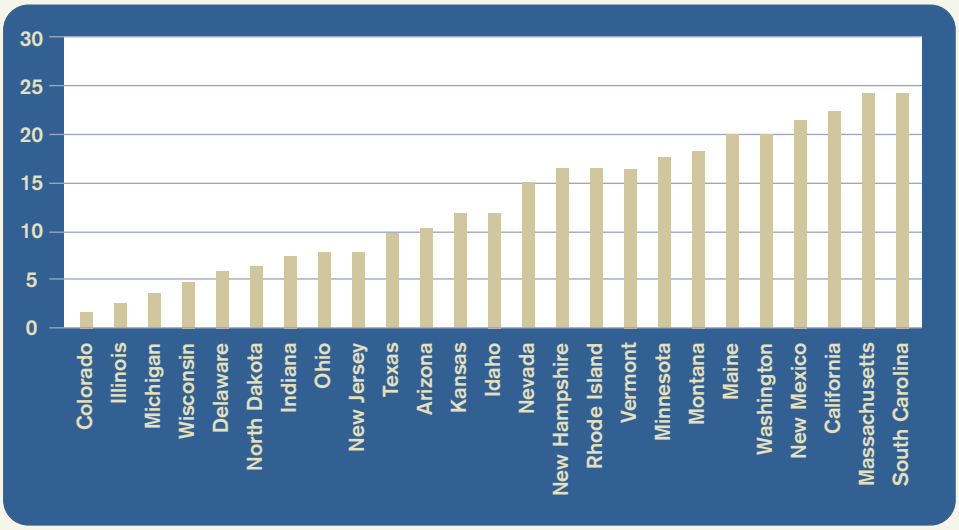
Note: This figure ranks the grade 8 math cut scores from easiest (Illinois) to most difficult (South Carolina) and shows the median difficulty across all states studied (in green).

Figure 5 – Average ranking of states according to the difficulty of their reading proficiency cut scores across all grades (higher ranks = more difficult standards)



Note: This figure shows the average rank in reading across all grades measured within a state, where a high rank denoted a high proficiency cut score. Colorado's reading cut scores had the lowest average rank, while South Carolina's cut scores had the highest average rank.

Figure 6 – Average ranking of states according to the difficulty of their mathematics proficiency cut scores across all grades (higher average ranks = more difficult standards)



Note: This figure shows the average rank in math across all grades measured within a state, where a high rank denoted a high proficiency cut score. Colorado's math cut scores had the lowest average rank, while South Carolina's cut scores had the highest average rank.

Differences in state proficiency standards are reflected in rigor of the curriculum tested.

The differences in standards are not numerical artifacts. They represent real differences in expectations.

To illustrate this point, we selected five states to represent the range of proficiency cut scores used for grade 4 reading (Table 1). We extracted questions from the MAP item pool that were equivalent in difficulty to the proficiency cut score for each of these states. To make comparison easier, all these items focused on a single reading skill that is commonly required in all state standards: the ability to distinguish fact from opinion. Almost all reading curricula have introduced this concept prior to fourth grade. Using the exhibits below, we can compare what "proficiency" requires in five different states.

Table 1 – Grade 4 reading proficiency cut scores for five states

Ranking	State	NWEA Scale Score associated with proficient	Percentile Rank
25/26	Colorado	187	11
24/26	Wisconsin	191	16
13/26	North Dakota	199	29
3/26	California	204	43
1/26	Massachusetts	211	65

Note: Colorado currently reports the state's "partially proficient" level of academic performance on its state test as "proficient" for NCLB purposes, while using the higher "proficient" level for internal state evaluation purposes. In effect, Colorado has two standards: an easier standard for NCLB, and a harder standard for internal state use. For purposes of fairly comparing Colorado to other states, we used their NCLB-reported standard. Consequently, all subsequent references to "proficient" or "proficiency" in Colorado should be understood as referring to NCLB-reported standard.

Reading Exhibit 1 – Grade 4 item with difficulty equivalent to Colorado's proficiency cut score (scale score 187 – 11th percentile)

Alec saw Missy running down the street. Alec saw Paul running after Missy. Paul was yelling, "Missy, stop! Wait for me!"

What do we know for sure?

A. Missy is Paul's big sister, and she is mad at him.

B. Paul is mad at Missy and is chasing her down the street.

C. Alec saw Paul running after Missy and calling for her to wait.

D. Alec tried to stop Missy because Paul wanted to talk to her.

Almost all fourth-graders answer this item correctly. It contains a very simple passage and asks the student to identify the facts in the passage without making an inference. The student does not have to understand terms like "fact" or "opinion" to correctly answer the question.

Reading Exhibit 2 – Grade 4 item with difficulty equivalent to Wisconsin's proficiency cut score (scale score 191 – 16th percentile)

Which sentence tells a fact, not an opinion?

A. Cats are better than dogs.

B. Cats climb trees better than dogs.

C. Cats are prettier than dogs.

D. Cats have nicer fur than dogs.

This item is also quite easy for most fourth-graders and does not require reading a passage. It does introduce the concepts of fact and opinion, however, and some of the distinctions between fact and opinion are subtle. For example, some children may believe that the differences in cat and dog fur are fact.

Reading Exhibit 3 – Grade 4 item with difficulty equivalent to North Dakota's proficiency cut score (scale score 199 – 29th percentile)

Summer is great! I'm going to visit my uncle's ranch in July. I will be a really good rider by August. This will be the best vacation ever!

Which sentence is a statement of fact?

A. Summer is great!

B. I'm going to visit my uncle's ranch in July.

C. I will be a really good rider by August.

D. This will be the best vacation ever!

Most fourth-graders answer this item correctly. The differences between fact and opinion in this item are considerably more subtle than in the prior item. For example, many fourth-graders are likely to believe that "Summer is great!" is not a matter of opinion.

Reading Exhibit 4 – Grade 4 item with difficulty equivalent to California's proficiency cut score (scale score 204 – 43rd percentile)

The entertainment event of the year happens this Friday with the premiere of Grande O. Partie's spectacular film Bonzo in the White House. This movie will make you laugh and cry! The acting and directing are the best you'll see this year. Don't miss the opening night of this landmark film—Bonzo in the White House. It will be a classic.

What is a fact about this movie?

A. It is the best film of the year.

B. You have to see it Friday.

C. It opens this Friday.

D. It has better actors than any other movie.

Just over half of fourth-graders from the MAP norm group answer this item correctly. The question requires the student to navigate a longer passage with more sophisticated vocabulary. Indeed, the student has to know or infer the meaning of "premiere" to answer the question correctly.

Reading Exhibit 5 – Grade 4 item with difficulty equivalent to Massachusetts's proficiency cut score (scale score 211 – 65th percentile)

Read the excerpt from "How Much Land Does a Man Need?" by Leo Tolstoy.

So Pahom was well contented, and everything would have been right if the neighboring peasants would only not have trespassed on his wheatfields and meadows. He appealed to them most civilly, but they still went on: now the herdsmen would let the village cows stray into his meadows, then horses from the night pasture would get among his corn. Pahom turned them out again and again, and forgave their owners, and for a long time he forbore to prosecute anyone. But at last he lost patience and complained to the District Court.

What is a fact from this passage?

A. Pahom owns a vast amount of land.

B. The peasant's intentions are evil.

C. Pahom is a wealthy man.

D. Pahom complained to the District Court.

This item is clearly the most challenging to read (it is Tolstoy after all), and the majority of fourth-graders in the NWEA norm group got it wrong. The passage is long relative to the others and contains very sophisticated vocabulary. At least three of the options identify potential facts in the passage that have to be evaluated.

ANALYSIS

When viewed in terms of items that reflect the difficulty of the five state standards, the differences in expectations are striking. The vocabulary used in the more difficult items is far more sophisticated than that used in the easier items. Moreover, students must be very careful in their analysis of the more difficult items to answer them correctly. Most compelling, however, are the sheer differences in the difficulty of the reading passages associated with these items, which range from something that could be found in a second-grade reader to a passage from Tolstoy.

For mathematics, we extracted examples of items with difficulty ratings equivalent to five states' proficiency cut scores in algebraic concepts (Table 2). None of the items requires computational abilities that would be beyond the scope of a typical grade 4 curriculum.

Table 2 – Grade 4 mathematics proficiency cut scores for five states

Ranking	State	NWEA Scale Score associated with proficient	Percentile Rank
25/25	Colorado	191	8
23/25	Illinois	197	15
13/25	Texas	205	34
3/25	California	212	55
1/25	Massachusetts	220	77

Math Exhibit 1 – Grade 4 math item with difficulty equivalent to Colorado's proficiency cut score (scale score 191 – 8th percentile rank)

Tina had some marbles. David gave her 5 more marbles. Now Tina has 15 marbles. How many marbles were in Tina's bag at first?

What is this problem asking?

A. How many marbles does Tina have now?

B. How many marbles did David give to Tina?

C. Where did Tina get the marbles?

D. How many marbles was Tina holding before David came along?

E. How many marbles do Tina and David have together?

Math Exhibit 1 shows an item that reflects the Colorado NCLB proficiency cut score. It is easily answered by most fourth-graders. It requires that students understand the basic concept of addition and find the right question to answer, although students need not actually solve the problem.

Math Exhibit 2 – Grade 4 math item with difficulty equivalent to Illinois' proficiency cut score (scale score 197- 15th percentile)

Marissa has 3 pieces of candy. Mark gives her some more candy. Now she has 8 pieces of candy. Marissa wants to know how many pieces of candy Mark gave her.

Which number sentence would she use?

A. 3 + 8 = ?

B. 3 + ? = 8

C. ? X 3 = 8

D. 8 + ? = 3

E. ? – 3 = 8

This item, reflecting the Illinois cut score, is slightly more demanding but is also easily answered by most fourth-graders. It requires the student to go beyond understanding the question to setting up the solution to a one-step addition problem.

Math Exhibit 3 – Grade 4 math item with difficulty equivalent to Texas's proficiency cut score (scale score 205 - 34th percentile)

Chia has a collection of seashells. She wants to put her 117 shells into storage boxes. If each storage box holds 9 shells, how many boxes will she use?

Which equation best represents how to solve this problem?

A. 9 – 117 = ?

B. 9 ÷ 117 = ?

C. 117 X 9 = ?

D. 117 + 9 = ?

E. 117 ÷ 9 = ?

This item, at a difficulty level equivalent to the Texas cut score, is answered correctly by most fourth-graders but is harder than the previous two. The student not only must be able to set up the solution to a simple problem, but must also know how to frame a division problem in order to answer the question correctly.

Math Exhibit 4 – Grade 4 math item with difficulty equivalent to California's proficiency cut score (scale score 212 - 55th percentile)

8 + 9 = 10 + ?

A. 6

B. 9

C. 17

D. 7

E. 6

Most fourth-grade students in the MAP norm group do not answer this question correctly. The more advanced concept of balance or equivalency within an equation is introduced in this item. This concept is fundamental to algebra and makes this much more than a simple arithmetic problem. The student must know how to solve a problem by balancing the equation.

Math Exhibit 5 – Grade 4 math item with difficulty equivalent to Massachusetts's proficiency cut score (scale score 220 - 77th percentile)

The rocket car was already going 190 miles per hour when the timer started his watch. How fast, in miles per hour, was the rocket car going seven minutes later if it increased its speed by 15 miles per hour every minute?

A. 205
D. 1330
B. 295
E. 2850
C. 900

This is obviously the most demanding item of the set and is not answered correctly by most fourth-graders within the MAP norm group. The student must understand how to set up a multiplication problem using either a one-step equation – ***190 + (7 x 15) = ?***—or a multi-step equation—***190 + (15+15+15+15+15+15+15) = ?***

ANALYSIS

These examples from reading and mathematics make it apparent that the states we studied lack a shared concept of proficiency. Indeed, their expectations are so diverse that they risk undermining a core objective of NCLB—to advance educational equality by ensuring that all students achieve their states' proficiency expectations. When the proficiency expectations in grade 4 mathematics range from setting up simple addition problems to solving complex, multi-step multiplication problems, then meeting these expectations achieves no real equity. The reading examples, too, show that "proficiency" by no means indicates educational equality. A student who can navigate the California or Massachusetts reading requirements has clearly achieved a much different level of competence than has one who just meets the Colorado or Wisconsin proficiency standard.

The proficiency expectations have a profound effect on the delivery of instruction in many states. Because of the consequences associated with failure to make adequate yearly progress (AYP), there is evidence that instruction in many classrooms and schools is geared toward ensuring that students who perform near the proficiency bar pass the state test (Neal and Whitmore-Schanzenback 2007). In Illinois, for example, this is apt to mean that some classrooms will place greater emphasis on understanding simple math problems like the one in Math Exhibit 2, while California and Massachusetts students are working with algebraic concepts of much greater sophistication, such as those in Math Exhibits 4 and 5.

Standards for mathematics are generally more difficult to meet than those for reading.

Figures 7 and 8 compare the proficiency cut score estimates for grades 3 and 8 in reading and mathematics. They show that in third grade, the mathematics standards are more difficult for students than are the reading standards in fourteen of the twenty-five states studied, while in eighth-grade the math standards are more difficult in twenty of the twenty-two states (eighth-grade math estimates were unavailable in three states).

ANALYSIS

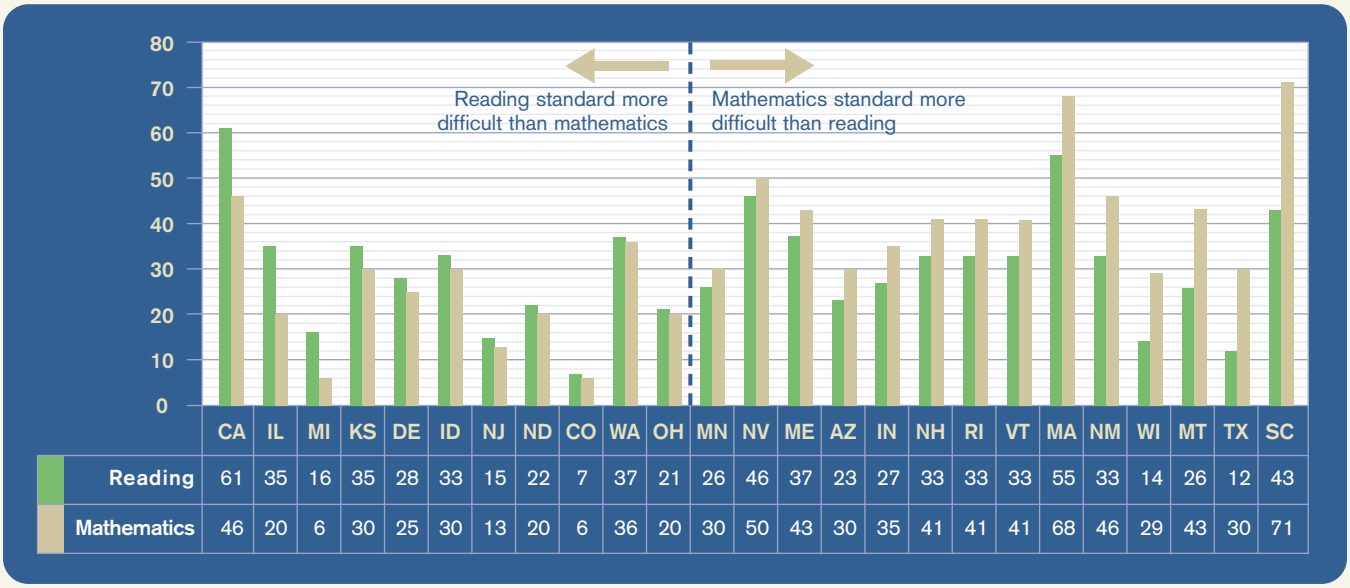
This interesting phenomenon may suggest that those who have argued for higher mathematics standards have effectively advanced their case. Of course, it also raises some questions. For example, if math skills are important enough to warrant setting a proficiency cut score at about the 67th percentile for Massachusetts eighth-graders, are reading skills so much less important that a cut score at the 31st percentile can be justified?

When the reading and mathematics proficiency standards differ greatly in difficulty, it can create confusion among policymakers, parents, the public, and educators, who may assume that proficiency represents a consistent standard of performance across subjects. Such consistency was not the case in many of the states examined in the current study, and the resulting discrepancies in proficiency expectations can make it difficult to judge the effectiveness of schools.

To further illustrate the discrepancy between math and reading standards, consider the differences in reported proficiency rates between reading and mathematics in Massachusetts. Figure 9 shows the state-reported proficiency rates by grade for reading and mathematics in 2006. These data show that 74 percent of students achieved the eighth-grade reading standard, while only 40 percent achieved the eighth-grade math standard.

Given only the information displayed in Figure 9, one might well conclude that Massachusetts schools have been much more effective at teaching reading than math. Yet when one examines the differences in the difficulty of the reading and mathematics cut scores at each grade (Figure 10), an entirely different picture emerges. In every grade, the proficiency cut score in mathematics is far more difficult than that in reading. (This is especially true by eighth grade, where the difference in cut scores is so large that, among the norm group, nearly twice as many students would pass reading than mathematics. As reported earlier, Massachusetts's third-grade reading cut scores are among the highest in the nation.) Thus, the state-reported differences in achievement are more likely a product of differences in the difficulty of the cut scores than differences in how well reading and math are taught.

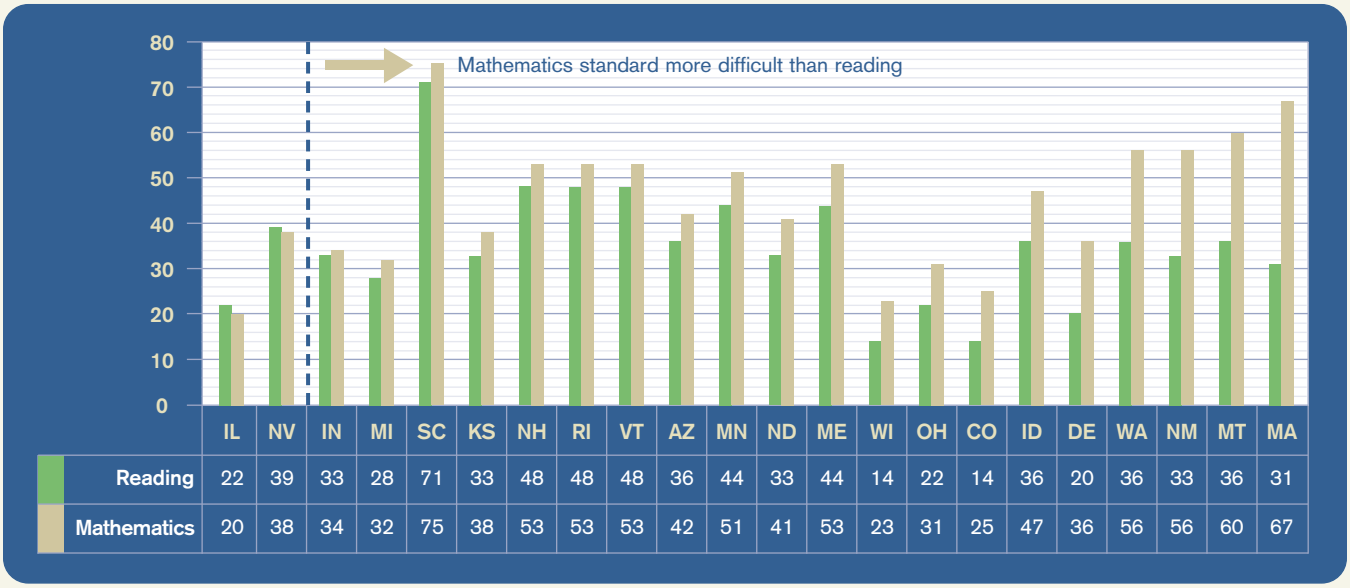
Figure 7 - Grade 3 reading and mathematics proficiency estimates (ordered by size of difference as shown by MAP percentile)



	CA	IL	MI	KS	DE	ID	NJ	ND	CO	WA	OH	MN	NV	ME	AZ	IN	NH	RI	VT	MA	NM	WI	MT	TX	SC
Reading	61	35	16	35	28	33	15	22	7	37	21	26	46	37	23	27	33	33	33	55	33	14	26	12	43
Mathematics	46	20	6	30	25	30	13	20	6	36	20	30	50	43	30	35	41	41	41	68	46	29	43	30	71

This shows the differences in difficulty of the third-grade math and reading standards across states. In nine of twenty-five states, the reading cut scores are more difficult. In sixteen of twenty-five states, the math cut scores are more difficult.

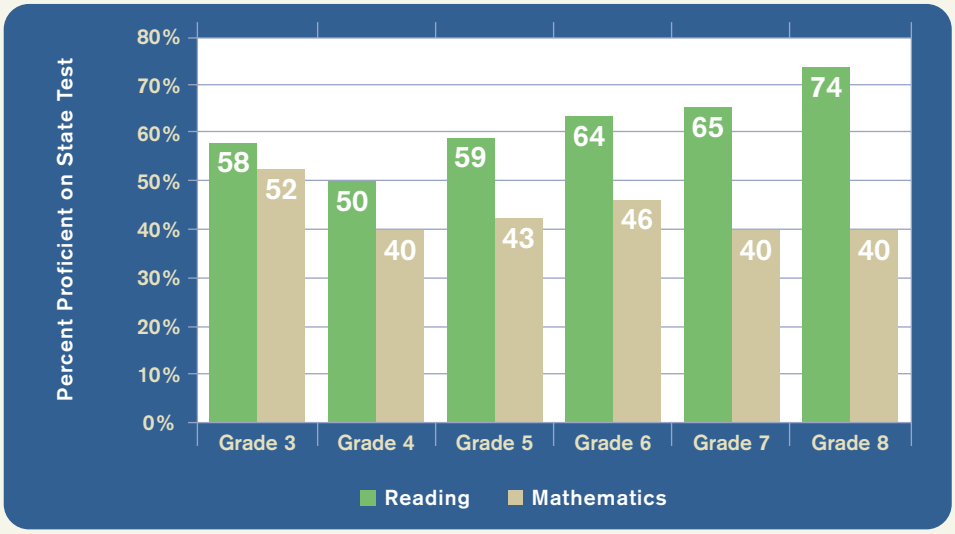
Figure 8 - Grade 8 reading and mathematics proficiency estimates (ordered by size of difference as shown by MAP percentile)



	IL	NV	IN	MI	SC	KS	NH	RI	VT	AZ	MN	ND	ME	WI	OH	CO	ID	DE	WA	NM	MT	MA
Reading	22	39	33	28	71	33	48	48	48	36	44	33	44	14	22	14	36	20	36	33	36	31
Mathematics	20	38	34	32	75	38	53	53	53	42	51	41	53	23	31	25	47	36	56	56	60	67

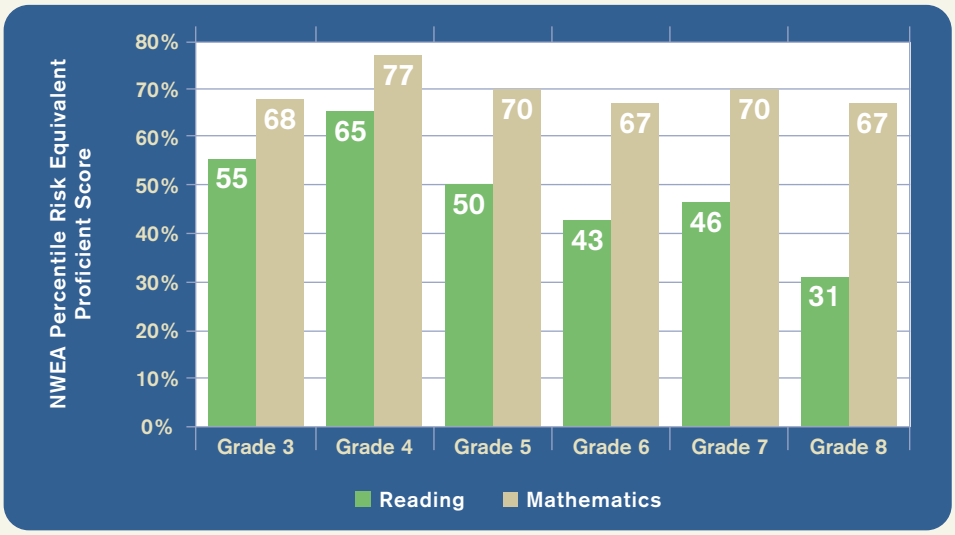
This figure shows the differences in difficulty of the eighth-grade math and reading standards across states. Math cut scores were more difficult than reading in twenty of the twenty-two states for which eighth-grade reading and math scores were estimated.

Figure 9 – State-reported proficiency rates in reading and mathematics, 2006 – Massachusetts



Note: This figure shows that at a higher percentage of students met the standards for reading proficiency than math proficiency at each grade.

Figure 10 – Proficiency cut score estimates for reading and mathematics, 2006 – Massachusetts (ranked by MAP percentile)



Note: This figure shows that the proficiency cut score on the state test is more difficult in math than in reading at every grade.

Two sample items (Reading Exhibit 6 and Math Exhibit 6) illustrate the difference in difficulty between the reading and math standards.

Reading Exhibit 6 – Grade 8 item with difficulty equivalent to Massachusetts's proficiency cut score (scale score 216 – 31st percentile)

Read the passage.

Katya's eyes adjusted to the dimness. She could tell that someone had once inhabited this place. She noticed markings on the walls, and she knew they would be a significant part of her archaeological study. There were jagged lines of lightning and stick figures.

What story element has the author developed within this passage?

A. theme
C. conflict
B. plot
D. setting

This reading item has the same difficulty as the Massachusetts grade 8 reading cut score and is answered correctly by the vast majority of eighth-graders. The passage is not complex, and students who are familiar with the literary concept of setting will answer it correctly.

Math Exhibit 6 – Grade 8 math item with difficulty equivalent to Massachusetts' proficiency cut score (scale score 242 – 67th percentile)

Maria has $5.00 more than Joseph. Together they have $37.50. Which of these equations would you use to find the amount of money Joseph has?

A. $j + (5 \times j) = \$37.50$

B. $j + (j \div 5) = \$37.50$

C. $5 \times j = \$37.50 + j$

D. $2 \times (j + 5) = \$37.50$

E. $j + j + 5 = \$37.50$

This item has the same difficulty as the Massachusetts mathematics proficiency standard and is missed by the majority of eighth-grade students in the NWEA norm group. The question is a multi-step problem and addresses a concept commonly found in Algebra I. Although the items in these two exhibits come from different disciplines, we know that the mathematics item is empirically more difficult than the reading item because far fewer eighth-graders within the NWEA norm group successfully answer the math item than the reading item.

ANALYSIS

In Massachusetts, the differences in the difficulty of the standards largely explain the differences in student performance. In eighth grade, 74 percent of Massachusetts pupils achieved the reading proficiency standard, while only 40 percent achieved proficiency in mathematics. A person viewing these data could easily come to several conclusions about curriculum and instruction in Massachusetts that would be erroneous. One could wrongly reach any of the following conclusions:

- Students perform more poorly in mathematics than in reading within Massachusetts.
- Educators teaching mathematics in Massachusetts are less competent than educators teaching reading in the state.
- The mathematics curriculum used for students in Massachusetts is not pushing the students as hard as the reading curriculum, thus resulting in poorer outcomes.
- Less instructional time is devoted to teaching math in Massachusetts than reading, thus resulting in poorer outcomes.

However, the truth is that students in the NWEA norm group would have produced the same disparity in achievement. In other words, had students from the multi-state NWEA norm group been compared to the same Massachusetts standards, a similar gap in achievement would have been found.

Experts sometimes assume that standard setting is a scientific process and thus that these sorts of differences in math and reading standards represent genuine differences in what is needed to be "proficient" in the real world. But as we have already shown, "proficient" is a concept that lacks any common definition. In truth, differences in reading and mathematics standards may emerge because of factors that have nothing to do with real-world requirements. For example, when states convene experts to set standards, they commonly select educators with high levels of competence in their field. In reading, the best-educated teachers commonly work with the lowest-performing readers, because those students require that kind of expertise. In mathematics, the opposite is typically true, with the best-educated instructors commonly teaching the most advanced courses. Thus differences in the makeup of the standard-setting group may well have more bearing on discrepant reading and mathematics expectations than do requirements for proficiency in the real world.

In any case, whether knowingly or not, many states have clearly set higher expectations for mathematics performance than they have for reading. Unfortunately, school systems and policymakers may infer from the resulting differences in performance that students in a given state have some deficiency in mathematics requiring special intervention. They may act on these kinds of inferences, allocating resources to address seeming gaps in math achievement that may not exist. As a consequence, resources might not be allocated to address problems with reading programs that remain hidden beneath this veneer of seemingly superior performance.

This is not to argue that math and reading standards must be equivalent in difficulty. One can defend different standards if the differences are intentional, quantified, and transparent. If educators and the public believe that math standards should be tougher than those in other subjects, if they understand that the mathematics standards will be more challenging to achieve, and if the state reports student performance with a transparency that ensures that the public will understand these differences, then discrepant standards can represent a rational and purposeful public policy choice. In reality, however, we rarely see the question of discrepant standards raised or addressed. This is regrettable, because in at least ten of the states we studied, there are wide differences in the difficulty of mathematics and reading standards that explain most of the difference in student achievement in those subjects.

Some might suggest that U.S. reading performance really is stronger than U.S. math performance and thus a reading standard set at, say, the 20th percentile (of a nation of good readers) is equivalent to a math standard set at, say, the 40th percentile (of a nation of children bad at math). We reject this hypothesis. It's true that international studies of student performance in reading and math have found that higher percentages of U.S. students achieve the top-level proficiency benchmarks in reading than achieve the top-level benchmarks in mathematics (Mullis, Martin, Gonzales, and Kenney 2003; Mullis, Martin, Gonzales, and Chrotowski 2004). Yet these studies examine math and reading performance separately, making no direct comparisons between the relative difficulties of the international math and reading benchmarks. Consequently, differences in math and reading performance in such studies are not directly comparable. Furthermore, as illustrated in the Massachusetts example above, any fair look at test items representative of the various standards would show real differences between math and reading expectations.

The purpose of the NCLB was to establish a common expectation for performance within states, presumably to ensure that schools address the learning needs of all children. Unfortunately, the disparity in standards between states undermines this purpose. While it may advance the cause of equity within Michigan to require all students to reach the 6th percentile in grade 3 mathematics, Michigan students are collectively disadvantaged when schools in most other states pursue far more challenging proficiency standards—standards that would, if achieved, leave students in Kalamazoo far behind their peers in Fort Wayne, Indiana, or St. Cloud, Minnesota.

Indeed, the sometimes-immense gaps in the difficulty of standards from state to state hardly seem rational. A barely proficient student in Michigan in no way resembles a barely proficient student in Massachusetts, and, unfortunately, a proficient reader in Massachusetts has achieved a far less difficult standard than one who meets the state's mathematics expectations.

Question 2: Is there evidence that states' expectations for proficiency have changed over time? If so, are state proficiency cut scores becoming more or less difficult?

Proficiency cut score estimates were generated at two points in time for nineteen states. Table 3 shows the states and time periods—all subsequent to NCLB's enactment—for which these estimates were generated. It also indicates whether a state announced changes to its assessment system or its cut scores during the period between our two estimates and briefly describes any changes that were made.

Of the nineteen relevant states, eight revised their scales or adjusted their proficiency cut scores. Of these, five adopted new measurement scales, while the other three changed the cut score on their existing scale in at least one grade. The remaining eleven states announced no changes to their proficiency cut scores during the period of the study. Of these, six added testing in some grades but did not change their cut scores in the other grades.

Table 3 - Reported action on state cut scores, 2002-2006

State	First Estimate	Second Estimate	Did state cut score change?	Date	Comments
Arizona	Spring 02	Spring 05	Yes	Spring 05	The state added grades to the assessment and adopted a new scale.
California	Spring 03	Spring 06	No		The state maintained the same scale and announced no changes to proficiency cut scores.
Colorado	Spring 02	Spring 06	No		The state maintained the same scale and announced no changes to proficiency cut scores. The state added tests and established cut scores for mathematics in grades 3 and 4.
Delaware	Spring 05	Spring 06	Yes	Spring 06	The state added grades to the assessment. The state maintained the same scale but announced changes to the cut scores. Officials reported raising cut scores slightly in reading in grades 3, 5, and 8 and lowering them slightly in math in grades 5 and 8.
Idaho*	Spring 02	Spring 06	No		The state used NWEA tests and scale during the period studied. We did not estimate cut score changes for Idaho.
Illinois	Spring 03	Spring 06	Yes	Spring 06	The state maintained the same scale. The state established cut scores for new grades added (4, 6, 7). The state reported lowering the grade 8 math proficiency cut score.
Indiana	Fall 02	Fall 06	No		The state maintained the same scale and announced no changes to cut scores. However, cut scores for new grades were established (4, 5, 7).
Maryland	Spring 05	Spring 06	No		The state maintained the same scale and announced no changes to cut scores. The test was expanded to add new grades.
Michigan	Fall 03	Fall 05	Yes	Fall 05	The state expanded the test to include more grades and introduced a new scale.

Table 3 - continued

State	First Estimate	Second Estimate	Did state cut score change?	Date	Comments
Minnesota	Spring 03	Spring 06	Yes	Spring 06	The state expanded the test to include more grades and introduced a new scale.
Montana	Spring 04	Spring 06	No		The state added grades but maintained the same scale and announced no changes to proficiency cut scores during the period of the study.
Nevada	Spring 03	Spring 06	No		The state maintained the same scale and announced no changes to proficiency cut scores. The test was expanded to include more grades.
New Hampshire	Fall 03	Fall 05	Yes	Fall 05	The state changed from its own assessment to the New England Common Assessment Program in 2005. The grades tested were expanded and a new scale was introduced.
New Jersey	Spring 05	Spring 06	No		The state maintained the same scale and announced no changes to proficiency cut scores during the period of the study. The state implemented the NJ ASK assessment in 2003 and included more grades in the assessment in 2006.
New Mexico	Spring 05	Spring 06	No		The state maintained the same scale and announced no changes to proficiency cut scores during the period of the study. The state changed to the current New Mexico Student Based Assessment in spring 2004.
North Dakota	Fall 04	Fall 05	No		The state maintained the same scale and announced no changes to proficiency cut scores during the period of the study.
South Carolina	Spring 02	Spring 06	No		The state maintained the same scale and announced no changes to proficiency cut scores throughout the study period.
Texas	Spring 03	Spring 06	Yes	Spring 03	The state maintained the same scale during the study period. Initial cut scores were established in spring 2003. According to the state, higher proficiency cut scores were phased in over a three-year period.
Washington	Spring 04	Spring 06	No		The state maintained the same scale and announced no changes to cut scores during the period of the study.
Wisconsin	Fall 03	Fall 06	Yes	Fall 05	The state implemented a new scale in fall 2005 and set new proficiency cut scores. The state reported using methods to try to maintain stability in the difficulty of the cut scores throughout the study period.

Table 3 outlines the official adjustments made by states to their proficiency cut scores. For the nineteen states in this part of the study, we were able to estimate cut scores at two points in time in sixty-four instances in reading and fifty-six instances in mathematics across grades 3 through 8. Any instance in which the estimated cut score changed by three or more scale score points was defined for purposes of this study as a substantive change in the mapped cut score. Three scale score points was used because it represents the typical student's standard error of measurement on the MAP assessment. Here's what we found.

Most state tests have not changed in difficulty in recent years. Changes that were observed were more often in the direction of less difficulty than of greater. The greatest declines in difficulty were in states with the highest standards.

Tables 4 and 5 summarize the direction of estimated changes by state and grade level for each subject. In reading, cut score estimates declined in two or more grades in seven states: Arizona, California, Colorado, Illinois, Maryland, Montana, and South Carolina. Among these states, only Arizona and Illinois changed their cut scores during the period studied. Reading cut score estimates increased in at least two grades in Texas and New Hampshire, both states that introduced changes to their tests or cut scores between the periods estimated, as well as in New Jersey, which did not introduce changes. In mathematics, cut score estimates declined in two or more grades in six states (Arizona, California, Colorado, Illinois, New Mexico, and South Carolina) and increased in two or more grades in Minnesota, New Hampshire[1], and Texas. Thus, eight states saw their reading and/or math tests become significantly easier in at least two grade levels, versus four states whose tests became harder.

[1] New Hampshire used the "basic" performance level to report Adequate Yearly Progress prior to joining the NECAP. Since adopting NECAP, the state reports the test's "proficient" level for purposes of AYP.

Table 4 – Directions of changes in reading proficiency cut score estimates by state and grade level

State	Estimates	Change	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Arizona	2002 & 2005	New Scale	→		↓			↓
California	2003 & 2006	None	→	↓	→	→	↓	↓
Colorado	2002 & 2006	None	↓	↓	↓	→	→	→
Delaware	2005 & 2006	Changed Cut Scores						→
Illinois	2003 & 2005	Changed Cut Scores	↓		→			↓
Indiana	2002 & 2006	None	→			→		→
Maryland	2005 & 2006	None	↓	→	↓			
Michigan	2003 & 2005	New Scale		→			↓	
Minnesota	2003 & 2006	New Scale	↓		→			↑
Montana	2004 & 2006	None		↓				↓
Nevada	2004 & 2005	None	↓		→			
New Hampshire	2003 & 2005	New Scale	↑			↑		
New Jersey	2005 & 2006	None	↑	↑				
New Mexico	2005 & 2006	None	→	→	→	→	→	→
North Dakota	2003 & 2006	None	↓	→	→	→	→	→
South Carolina	2002 & 2006	None	↓	↓	↓	→	→	→
Texas	2003 & 2006	Changed Cut Scores	↑		↑	↑	↑	
Washington	2004 & 2006	None		↓			→	
Wisconsin	2003 & 2006	New Scale		→				↓

Table 5 – Direction of changes in mathematics-proficiency cut score estimates by state and grade level

State	Estimates	Change	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Arizona	2002 & 2005	New Scale	↓		↓			↓
California	2003 & 2006	None	→	→	↓	→	↓	
Colorado	2002 & 2006	None			↓	→	↓	↓
Delaware	2005 & 2006	Changed Cut Scores						↓
Illinois	2003 & 2005	Changed Cut Scores	→		↓			↓
Indiana	2002 & 2006	None	→			↓		→
Michigan	2003 & 2005	New Scale		↓				→
Minnesota	2003 & 2006	New Scale	→		↑			↑
Montana	2004 & 2006	None		↓				↑
North Dakota	2004 & 2005	None	→	→	↓	→	→	→
New Hampshire	2003 & 2005	New Scale	↑			↑		
New Jersey	2005 & 2006	None	↓	→				
New Mexico	2005 & 2006	None	→	→	→	↓	→	↓
Nevada	2003 & 2006	None	→		→			
South Carolina	2002 & 2006	None	→	→	→	↓	→	↓
Texas	2003 & 2006	Changed Cut Scores			↑		↑	
Washington	2004 & 2006	None		→			→	
Wisconsin	2003 & 2006	New Scale		→				↓

Note: Changes in tables 4 and 5 are depicted as increases (green arrow) or decreases (black arrow) when the difference in estimated cut scores is at least three scale score points (one student standard error of measurement). Changes of less than three points are represented by a blue arrow.

Figures 11 and 12 show the magnitude of changes in cut score estimates for each state and grade level. Although the majority of changes were not large enough to be considered substantive, the figures show that cut score estimates declined far more frequently than they increased. In reading, these changes were generally greatest in states that had the most difficult prior standards, while in math the changes were more even across the distribution. These figures also illustrate how changes in cut score estimates would affect the pass rate of students in the NWEA norming sample. Using South Carolina's grade 5 reading standard (SC5*) in Figure 12 as an example, the change in the estimated cut score lowered the difficulty of the reading proficiency standard from the 76th percentile to the 64th percentile. Thus if the our estimate of the current cut score were applied to the norming sample, we would estimate that 12 percent more students would pass South Carolina's test than would have passed in 2002, solely as a result in the change in our estimate of the difficulty of the standard, even if actual student achievement remained the same.

ANALYSIS

These trends do not indicate a helter-skelter "race to the bottom." They rather suggest more of a walk to the middle. The states with the greatest declines in estimated cut scores were those with very high standards. At the same time, some states with low standards saw their cut score estimates increase. Though many factors could explain these changes (see pp. 34-35), it is possible that these states are reacting to the 100 percent proficiency requirement of the No Child Left Behind Act.

Figure 11 – Summary of reading cut score estimates by state and grade level (from highest prior cut score estimate to lowest)

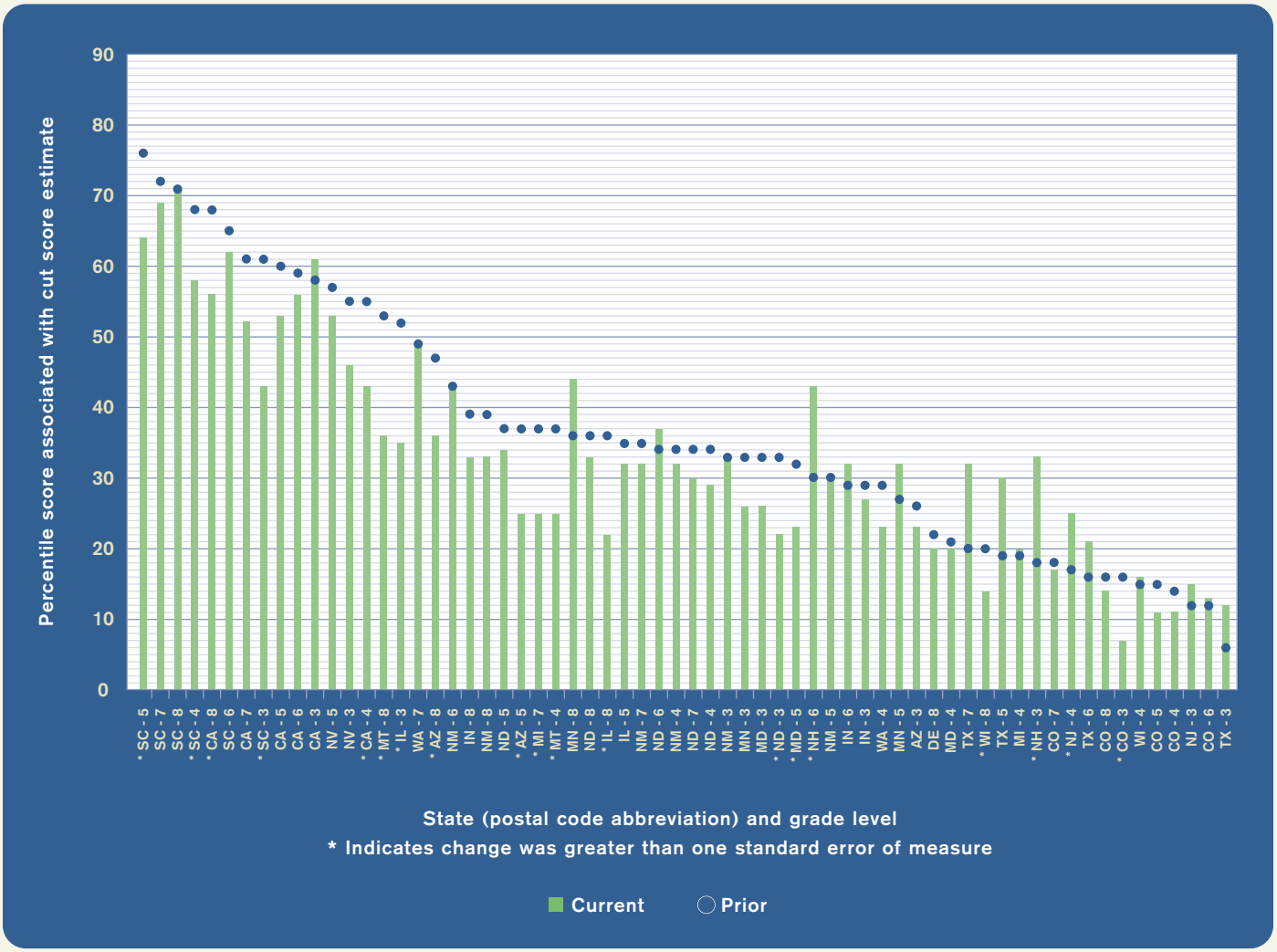


Figure 12 – Summary of mathematics cut score estimates by state and grade level (from highest prior cut score estimate to lowest)

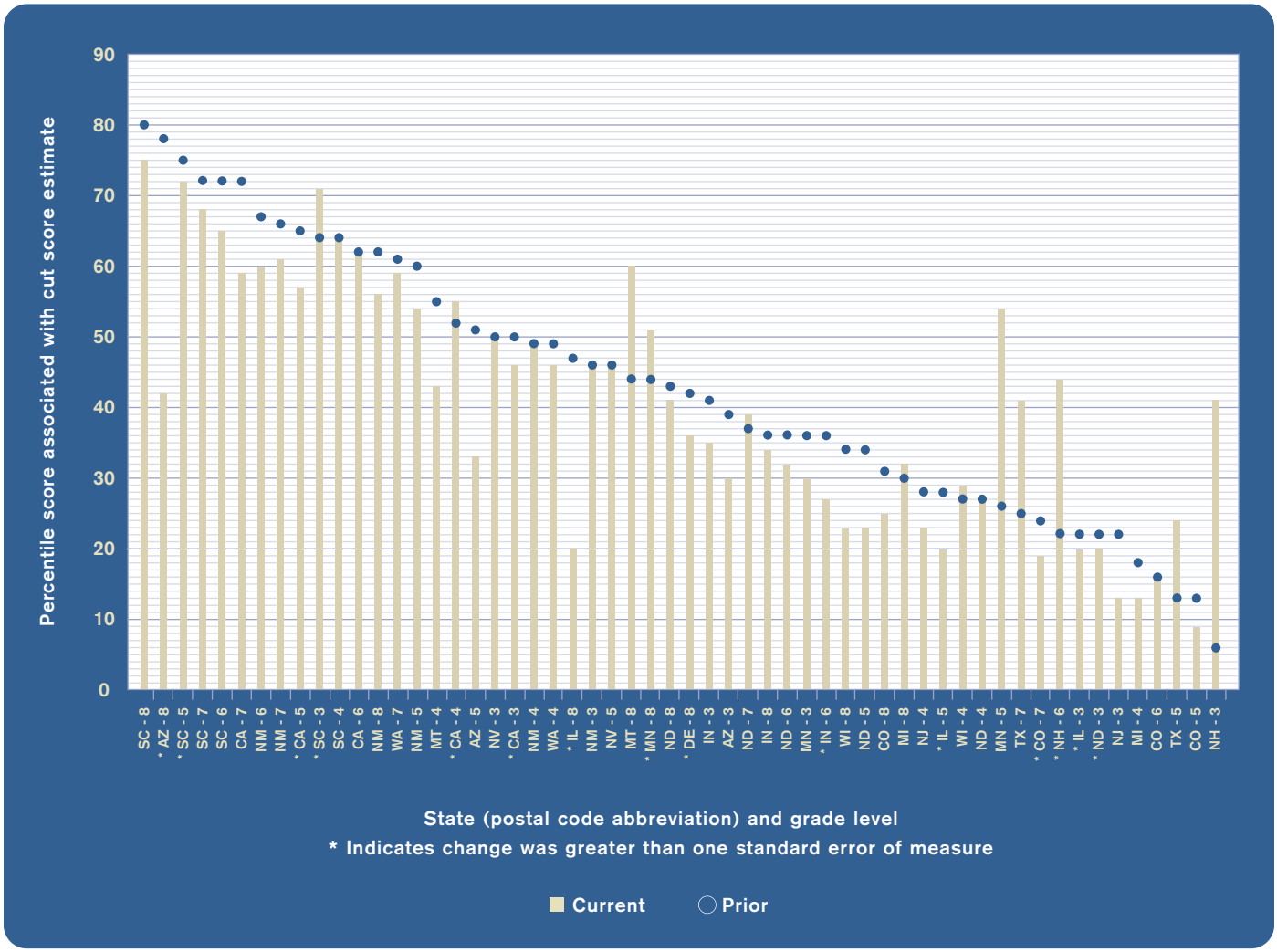


Table 6 – Summary of changes in proficiency cut score estimates

States that moved to new scale or officially changed cut scores				
	Increase	No Change	Decrease	Total
Reading	7 (35%)	6 (30%)	7 (35%)	20
Mathematics	6 (33%)	4 (22%)	8 (44%)	18

States that announced no changes to cut scores				
	Increase	No Change	Decrease	Total
Reading	2 (5%)	26 (59%)	16 (36%)	44
Mathematics	1 (3%)	24 (63%)	13 (34%)	38

Note: This table shows, for example, that among states that announced no changes to their reading cut scores, cut score estimates increased 5 percent of the time, decreased 36 percent of the time, and did not change 59 percent of the time.

We also disaggregated the data to differentiate between those states that made changes to their cut scores or adopted new measurement scales on the one hand, and those that announced no changes during the period studied on the other. Table 6 shows that among those states that announced changes, the number of increases in estimated cut scores roughly balanced with the number of declines. Among those states that announced no changes, however, more cut scores estimates declined than increased.

Changes in proficiency cut score estimates were inversely related to passing rates.

We evaluated the relationship between changes in our cut score estimates and passing rates on state proficiency tests. If changes in our cut score estimates have a strong inverse relationship to passing rates, that is, if passing rates improve when cut scores decline (based on NWEA estimates), then some portion of state-reported differences in passing rates can be explained by changes in test difficulty. If there is no correlation, then improvements in the state passing rate are more likely to reflect true improvements in student achievement that would be validated by other assessments. Put another way, if achievement goes up while the difficulty of the test remains the same, it lends credibility to the claim that achievement went up because students learned more.

Table 7 shows the correlation between our cut score estimates and the reported passing rates on state proficiency tests in reading and mathematics (the complete state-by-state data comparing cut score estimates and proficiency rates are available in Appendices 6 and 7). The results show strong inverse correlations between changes in cut scores and changes in state-reported proficiency rates, meaning that declines in proficiency cut score estimates were associated with increases in the state-reported proficiency rate, while increases in cut scores were associated with declines in the proficiency rate. In reading, the Pearson coefficient for all states and grade levels was -.71 with an r^2 of .50. This means that approximately 50 percent of the variance in the state proficiency rates could be explained by changes in the cut score. As expected, the correlation was slightly higher when the state made official changes to its cut score. In those cases, the Pearson r was -.79 with an r^2 of .63, meaning 63 percent of the variance in

student proficiency rates was explained by the changes that occurred in the cut score. Nevertheless, the correlation was also relatively strong among states that maintained their cut scores, with changes in our estimate explaining almost half of the variance in student proficiency rates ($r = -.70$, $r^2 = .49$). Once again this would suggest that about half of the improvement in student performance in these states was explained by decreased difficulty in their tests.

In mathematics, a very strong inverse correlation ($r = -.84$) was found between changes in cut scores and changes in the state-reported proficiency rates for the entire group. Thus cut score changes would explain about 70 percent of the variation among state-reported proficiency rates ($r^2=.70$). Among those states that maintained their cut scores, however, the inverse correlation was only moderate ($r=-.56$), although still large enough to explain about 32 percent of the variation in cut scores.

Table 7 – Correlation between reading and mathematics cut score estimates and state-reported proficiency rates

READING

	N	Average cut score estimate change (in percentile ranks)	Average proficiency rate change	Pearson r	R^2
All cases*	63	-3.30	2.47%	-0.71	0.50
State changed cut score*	19	-0.42	2.97%	-0.79	0.63
State did not change cut score*	44	-4.55	2.25%	-0.70	0.49

MATHEMATICS

	N	Average cut score estimate change (in percentile ranks)	Average proficiency rate change	Pearson r	R^2
All cases*	55	-2.20	4.38%	-0.84	0.70
State changed cut score*	17	0.06	5.83%	-0.93	0.87
State did not change cut score*	38	-3.21	3.73%	-0.56	0.32

* Delaware could not be included in this portion of the analysis because the state does not report proficiency percentages by grade.

ANALYSIS

These findings suggest that the primary factor explaining apparent gains in student proficiency rates is changes in cut score estimates. In terms of the improvement in student achievement that occurred between points at which the two estimates were made, half of the improvement in reading, and 70 percent of the improvement in mathematics, is probably idiosyncratic to the state test, and would not necessarily transfer to other achievement tests in these subjects.

In those cases in which the state did not adopt changes to its cut scores, what could cause our estimate to change? Because the NWEA scale is stable over time, the empirical explanation would be that student performance apparently changed on the state test without the same change in performance showing up on the NWEA assessment. Thus, some of the learning gains reported by state tests may be illusory. Several factors, most of which don't imply changes to the state test itself, but to the conditions and context surrounding it, could explain this phenomenon:

1. Educational Triage Strategies. Evidence is emerging that the accountability metrics used for No Child Left Behind may encourage schools to focus their improvement efforts on the relatively small numbers of students who perform near the proficiency bar on the state test. This triage strategy favors those students who can most help the school meet AYP requirements (Booher-Jennings 2005; White and Rosenbaum 2007; Neal and Whitmore-Schanzenbach 2007). If triage strategies were employed—and assuming they were effective—they would cause improvement in proficiency rates without parallel improvements in MAP, thus reducing our estimate of the cut score. For the majority of students who perform well above or below the proficiency bar, however, these strategies are not likely to improve learning.

2. Change in stakes. As NCLB's requirements are implemented, the consequences of poor performance on state tests have risen considerably for schools. Several prior studies have found strong relationships between the gains in student achievement and the implementation of high-stakes testing (Carnoy and Loeb 2002; Rosenshine 2003; Braun 2004). Cronin (2006), however, found that student performance gains on the Idaho state test were largely explained by a reduction in the number of students who did not try on the test (i.e., they "tanked" it), relative to a comparison group of students taking a low-stakes test. It is possible therefore, that the stakes associated with state tests may increase the motivation of students taking the state test, without resulting in improvements in achievement that become visible on other assessments. If that were the case in this study, such a change would cause the cut scores estimated by the benchmark test (i.e., MAP) to decline.

3. Test preparation strategies. Teachers and students have access to a number of materials that help them prepare for their state test. This includes test blueprints, sample items, and, in a few states, entire copies of past state tests. Some publishers offer resources to help prepare students for these exams, and teachers may teach to the test—that is, focus instruction on particular content and skills that are

likely to be seen on their state test. Koretz (2005) and Jacob (2002) found declines in test scores when some change in the form of the standardized test rendered these particular strategies less useful. These kinds of test-preparation strategies would raise scores on a particular test without generalizing to the larger domain and would cause estimated cut scores on a companion test to decline.

4. Differences in test alignment. A state's tests are supposed to be carefully aligned to state academic standards so that they sample students' success in acquiring the skills and knowledge that the state believes students should have. Certain exams, such as the NAEP, are not necessarily aligned to the same standards. As we explained in the introduction, however, the MAP test is purposely aligned to each state's standards, so that this problem is minimized for this study. Nevertheless, there is content on some reading or English/language arts and on some mathematics tests that cannot be assessed using MAP; most obviously, for instance, MAP does not assess writing. Particularly in those states that combine reading with language arts testing, improvements in student writing performance would produce gains on the state test that would not be matched on MAP, and this could cause the MAP estimate of the cut score to decline. In addition, over time educators may have tightened the alignment of instruction to the state test in a manner that might keep improvements from being visible on other instruments.

5. Drift in the difficulty of the state test. The state test might have become less difficult over time without anyone intending it. One of the greatest challenges that psychometricians face is maintaining a constant level of difficulty in a test from year to year. Over time, despite earnest efforts, the difficulty of a scale may drift. This risk increases when a test has been in use for many years. If drift in the measurement scale causes one test to become easier relative to its companion test, estimated cut scores on the companion test would decline.

It's impossible to know which of these factors, if any, explains why our estimates of state cut scores declined. Regardless, they all leave doubt as to whether improved performance on state tests is real—whether, that is, it reflects true improvements in learning. This doubt could remain even if the state offered the identical test in 2006 as in 2003. Several prior studies have reached this same conclusion, finding that improvements in student performance on state tests have not paralleled results on other tests of the same domain (Triplett 1995; Williams, Rosa, McLeod, Thissen, and Stanford 1998; McGlaughlin 1998a, 1998b; Education Trust 2004; Cronin, Kingsbury, McCall, and Bowe 2005). The most recent, a study of state proficiency improvements relative to NAEP, found that learning improvements on state tests were not reflected in NAEP, and that changes in state testing programs were the likely explanation for most improvements in proficiency (Fuller, Wright, Gesicki, and Kang 2007).

These findings lead us to advise caution in interpreting the gains reported on some state assessments, since these gains may not in fact reflect robust improvements in student achievement of a kind that can be replicated by other tests or in other venues.

Question 3: How closely are proficiency standards calibrated across grades? Are the standards in earlier grades equal in difficulty to proficiency standards in later grades?

Standards are calibrated when their relative difficulty remains constant from grade to grade. In other words, mastery of the eighth-grade standard would pose the same challenge to the typical eighth-grader that mastery of the third-grade standard would pose for the typical third-grader. To illustrate, assume that the athletic proficiency standard for an eighth-grader performing the high jump is four feet. Let's assume further that 40 percent of eighth-graders nationally can jump this high. What should the standard at third grade be? If the standard is to be calibrated, it would be the height that 40 percent of third-graders could jump successfully—say, two feet. Consequently, a third-grader who can high-jump two feet can fairly be said to be on track to meet the eighth-grade standard.

Some have suggested that calibration undermines the purpose of standards, because the process establishes proficiency benchmarks by using normative information (how the students performed relative to each other) rather than criterion-based information (how the students performed relative to the expectations for the grade). But arguing for calibrated standards is not tantamount to arguing for normative standards. We maintain that standards should be criterion based at the end points of the educational process. In this case, we believe that the criteria for eighth-grade proficiency should be based on proper academic expectations for students completing middle school. Once these are known and clear, the standards for the prior grades should be empirically benchmarked so that one can say fairly and with reasonable accuracy that children attaining the state's standard at grade 3 are on track to meet the standard in grade 8.

One way to establish these benchmarks is to use a normative projection. To illustrate, assume we have a single scale that measures performance in reading across grades. Assume that the eighth-grade reading proficiency standard is set at a scale score of 250 points and let's further assume that 50% of eighth-graders meet or exceed this score. A third-grader would be considered to be on track for this standard if he or she performs at the 50th percentile of the group in the third-grade.

Another way to establish benchmarks is by using longitudinal student-growth information to project performance. Assume once again that the eighth-grade standard remains at a scale score of 250 points. Let's also assume that we have empirically demonstrated that historically, students who meet this cut score typically grew 30 points between fifth and eighth grades. If so, then a score of 220 would represent a calibrated benchmark standard for fifth grade, because students meeting this standard, assuming normal growth, would go on to meet the eighth-grade standard.

The process is somewhat akin to establishing benchmarks for a long trip. Someone wanting to travel from Portland, Oregon, to Chicago in four days—a 1,700-mile trip—needs to average 425 miles per day in order to arrive on time. Knowing that, the traveler also knows that she has to drive from Portland to Twin Falls, Idaho, on the first day to be on track and must reach Omaha, Nebraska, by the end of the third day to remain on track. If she doesn't meet these benchmarks, she will not make her destination on time unless she drives faster or longer to make up for the delays.

But the process mandated by NCLB is different. It in effect allows experts to set the destination for day 1 without first determining where exactly travelers would need to be at that point in order to reach the final destination at the intended time.[2]

It is important for standards to be calibrated. Ultimately, a third-grade educational standard does not exist for its own sake, but as a checkpoint or way station en route to a more important destination. Whether that ultimate destination is college readiness, work readiness, or high school proficiency, the purpose of intermediate attainment standards is to indicate whether students are on track to meet these goals. To extend the prior analogy, reaching the third-grade destination, i.e., proficiency in third grade, should provide some assurance to parents that their children will meet the eighth-grade standard if they keep "driving" their learning at the same rate. If standards aren't calibrated in this manner, we send

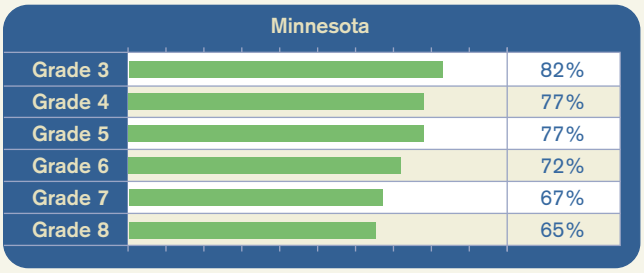
[2] The proficiency standards adopted in 2003 by the state of Idaho were developed using a process that calibrated the cut scores for grades 3 through 9 so they predicted success on the 10th-grade standard. This process was rejected by the U.S. Department of Education during peer review because the approach used did not account for "mastery of State content standards at specific grade levels" (United States Department of Education 2005).

confusing messages to educators, students, and families, who wonder why passing at one grade would not predict passing at another. Parents may blame the teacher or school for children's "poor performance" in their current grade when in truth the prior grade's standards were not challenging enough.

Reading and math tests in the upper grades are consistently more difficult to pass than those in earlier grades (even after taking into account obvious differences in student development and curriculum content).

The experience of Minnesota illustrates some of the issues that may be encountered when a proficiency standard is not calibrated across grades. Imagine that you are a parent viewing the results of the Minnesota Comprehensive Assessment – series II (MCAII) in the newspaper. Figure 13 shows the spring 2006 statewide reading results.

Figure 13 – Proportion of students scoring proficient or better on the Minnesota Comprehensive Assessment in reading (MCA-II), 2006



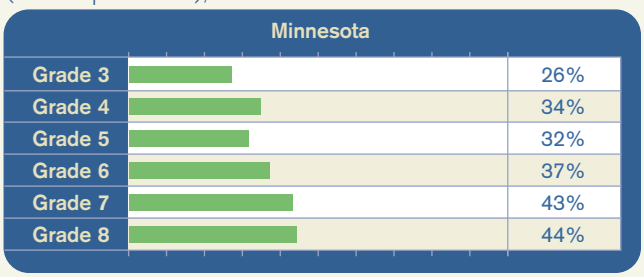
A parent interpreting these results would probably assume that third-graders in the state were doing far better than their peers in eighth grade. They might be concerned about the "deteriorating" performance in grades 7 and 8. Indeed, newspaper editorials, talk radio, and on-line discussions might identify a "crisis in the middle grades" and call for radical changes in the curriculum and organization of middle schools. Gradually, Minnesotans might come to believe that the discrepant results are a product of slumping middle school students and their lackluster teachers; meanwhile, they might believe that all is well in their elementary schools. Yet it is not clear that either inference would be warranted. If we look at Minnesota students' performance on the 2005 NAEP test in reading, shown in Table 8, we see that fourth- and eighth-graders perform about the same on their respective tests (albeit far below state-reported performance). Why then the grade-to-grade gap in performance on the Minnesota state assessment?

Table 8 – Minnesota performance on the 2005 NAEP in reading

	Grade 4	Grade 8
Percentage performing "proficient" or above	38%	37%

The answer lies in understanding that the difference in reported performance is really a function of differences in the difficulty of the cut scores and not actual differences in student performance. If we look at Figure 14, which shows the NWEA percentile ranks associated with the MCA-II proficiency cut scores for reading, we see that the third-grade cut score was estimated at the 26th percentile, meaning that 26 percent of the NWEA norm group would not pass a standard of this difficulty. By extension, 74 percent of NWEA's norm group *would* pass this standard. The proficiency cut score for eighth-grade, however, was estimated at the 44th percentile. This more difficult standard would be met by only 56 percent of the NWEA norm population.

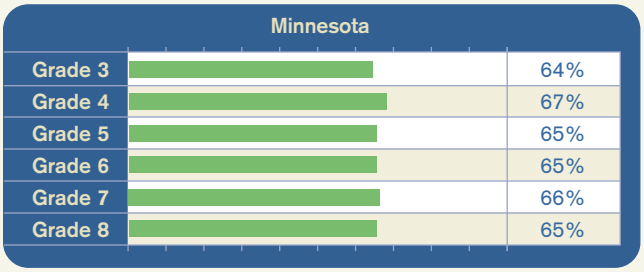
Figure 14 – Reading proficiency cut scores by grade (in MAP percentiles), 2006



Now we can see that the difference in reported performance reflects differences in the difficulty of the cut scores rather than any genuine differences in student performance. According to our estimates, because of the difference in difficulty of the standards, about 18 percent fewer eighth-graders would pass the Minnesota test in eighth-grade than passed in third (74% - 56% = 18%). And in fact the Minnesota results show that 17 percent fewer eighth-graders passed the MCA-II than third-graders.

What would happen if we adjusted the estimates of performance to reflect the differences in difficulty of the Minnesota proficiency standards, so that the proficiency cut score at each grade was equivalent to the eighth-grade difficulty level (Figure 15)?

Figure 15 – Estimated reading proficiency rate after calibrating to the 8th grade proficiency cut scores, 2006



The calibrated results indicate that there are no substantive grade-by-grade differences in reading performance. This is good news and bad news. The good news is that middle school students do not perform worse than their younger siblings in the earlier grades. The bad news is that we now know that far more third-, fourth-, and fifth-graders are at risk to miss the eighth-grade standards than we had previously believed. Using the data in Figure 14, a Minnesota student who performed at the 35th MAP percentile in reading in third-grade and maintained that percentile rank through eighth-grade would have been proficient in grades 3, 4, and 5 but not proficient in grades 6, 7, and 8.

Our analysis of proficiency standards found that in about 42 percent of the states studied, eighth-grade proficiency cut scores in reading were 10 percentile points or more difficult to achieve than the third-grade proficiency cut scores (Table 9).

In math, 68 percent of the states studied had eighth-grade proficiency cut scores that were 10 percentile points or more difficult to achieve than third-grade.

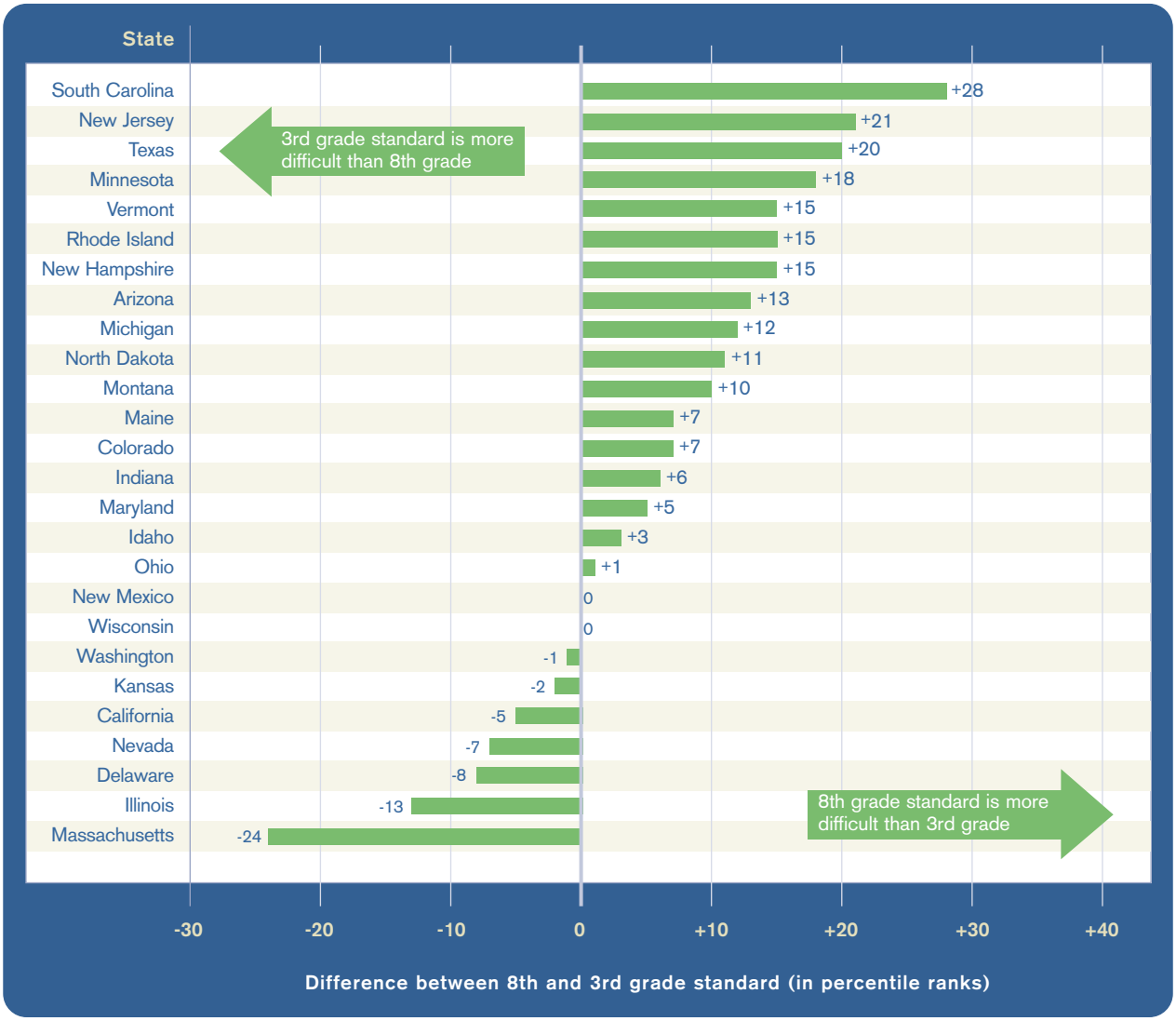
Figures 16 and 17 show the actual differences between the third- and eighth-grade proficiency cut scores for all of the states studied.

Table 9 – Differences between the difficulty of third- and eighth-grade proficiency standard*

	Reading	Mathematics
8th grade proficiency cut score was **somewhat** more difficult than 3rd grade (greater than 0 but less than 10 percentile ranks)	5/26 states (19%)	2/25 states (8%)
8th grade proficiency cut score was **substantially** more difficult than 3rd grade (by 10 or more percentile ranks)	11/26 states (42%)	17/25 states (68%)

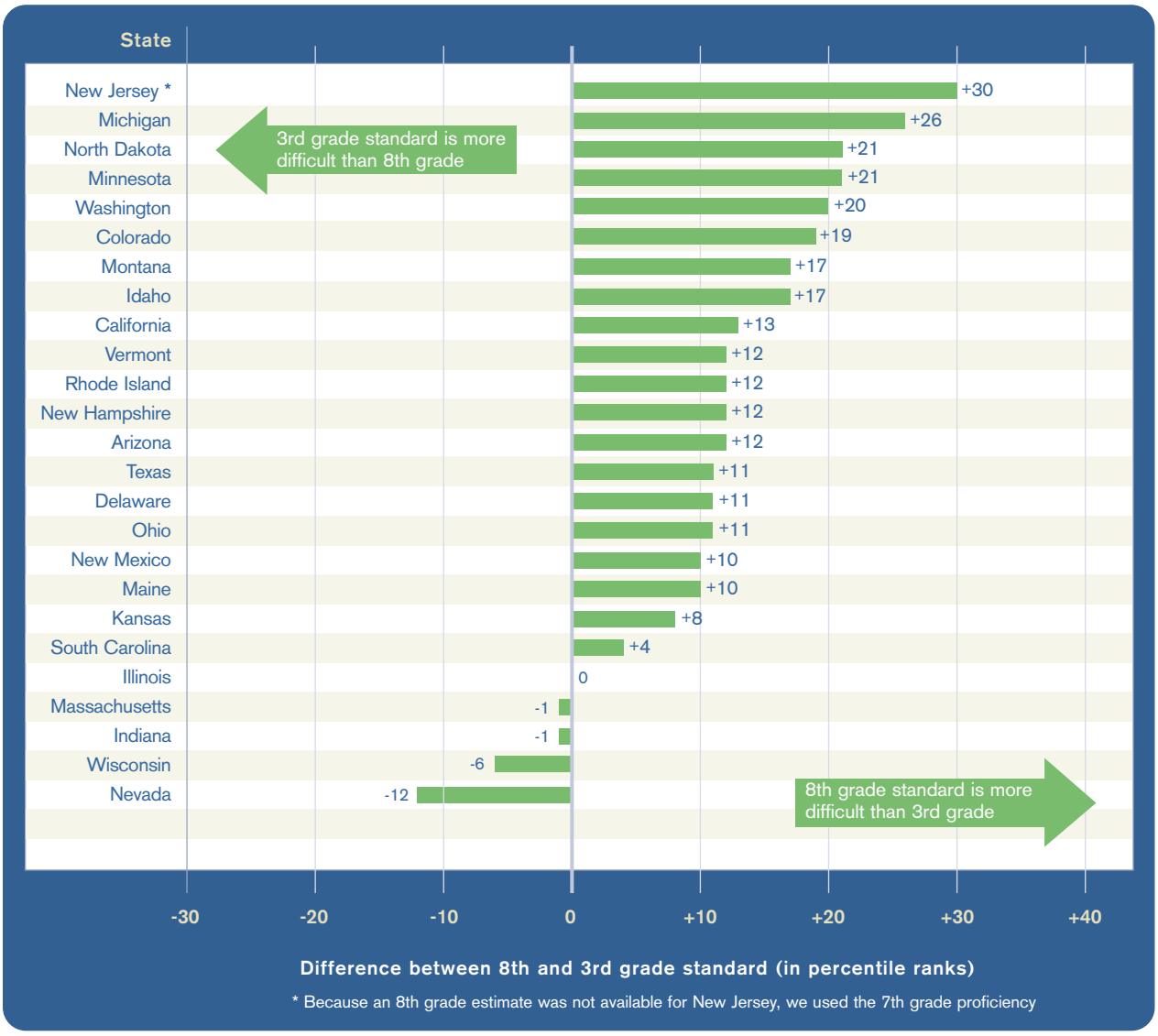
* Because 8th grade cut scores were not available, 7th grade proficiency cut scores were used in Texas for reading comparisons and in California, New Jersey, and Texas for mathematics comparisons

Figure 16 - Differences in third- and eighth-grade proficiency cut score estimates in reading (expressed in MAP percentiles)



Note: This figure shows, for example, that in Massachusetts, the third-grade reading standard is more difficult than the eighth-grade standard by 24 percentile points.

Figure 17 - Differences in third- and eighth-grade proficiency cut score estimates in mathematics (expressed in MAP percentiles)



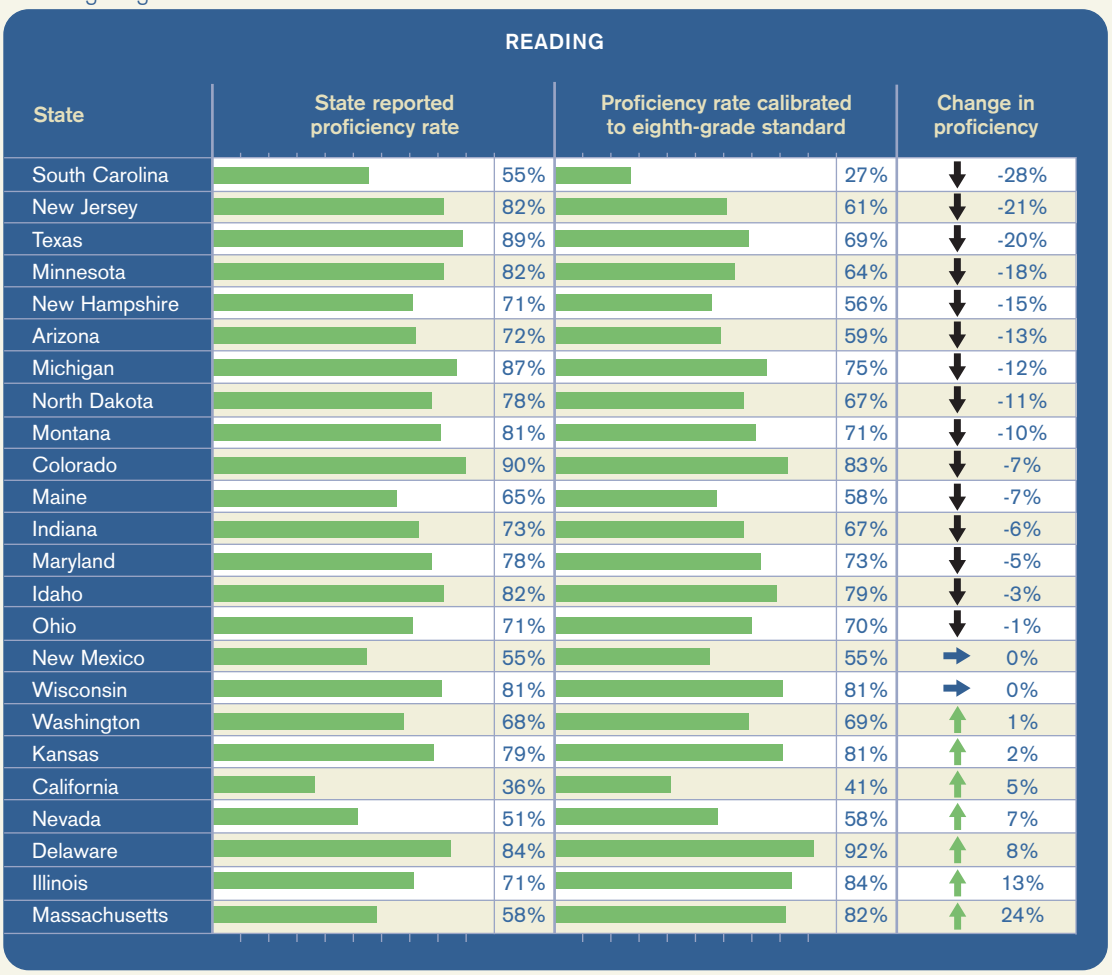
Figures 18 and 19 show how the current reported student proficiency rates for third grade might be affected if the third-grade standards were calibrated so that they were equivalent in difficulty to the eighth grade standards. In general, the data show that third-grade proficiency rates would decline, in some cases quite dramatically, if the third-grade standards reflected the performance level required for eighth-graders. In Texas, for example, we estimate that the third grade proficiency rate might be twenty points lower if the third grade reading test were calibrated to the difficulty of the eighth grade exam and that the third grade math results would be eleven points lower. Differences of similar magnitude in both reading and mathematics were found in many states, including Michigan, Minnesota, Montana, North Dakota, Texas, and the three states using NECAP (New Hampshire, Rhode Island, and Vermont).

ANALYSIS

These data make the problem obvious. Poorly calibrated standards create misleading perceptions about the performance of schools and children. They can lead parents, educators, and others to conclude that younger pupils are safely on track to meet standards when that is not the case. They can also lead policymakers to conclude that programs serving older students have failed because proficiency rates are lower for these students, when in reality, those students may be performing no worse than their younger peers. And conclusions of this sort can encourage unfortunate misallocations of resources. Younger students who might need help now if they are to reach more difficult standards in the upper grades do not get those resources because they have passed the state tests, while schools serving older students may make drastic changes in their instructional programs in an effort to fix deficiencies that may not actually exist.

Bringing coherence to the standards by setting initial standards that are calibrated to the same level of difficulty can help avoid these problems. If states begin with calibrated standards, then they know that between-grade differences in performance represent changes in the effectiveness of instruction, rather than in the difficulty of the standard. Armed with this knowledge, schools can make better use of resources to address weaknesses in their programs and can build on strengths.

Figure 18 – State-reported reading proficiency rates for third grade, before and after calibration to the eighth-grade standards.

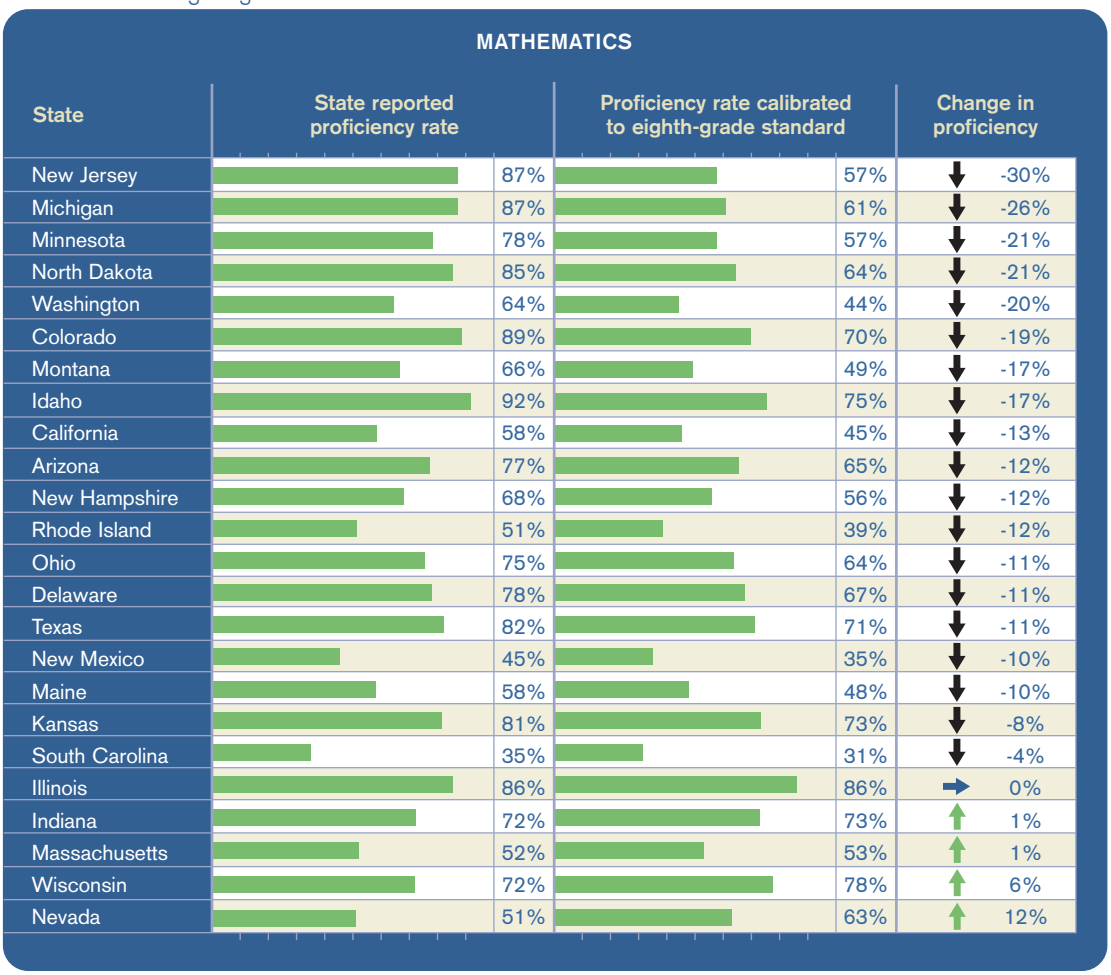


Discussion

It is essential to have high-quality educational standards. Properly implemented, such standards communicate the level at which a student must perform in order to meet their educational aspirations. Properly implemented, such standards are stable, so that stakeholders can evaluate whether students are making progress toward them over time. Properly implemented, such standards are calibrated across grades, so that, assuming normal growth, parents and students can have confidence that success at one grade level puts students on track for success at the completion of their education.

Unfortunately, the current system of standards is not properly implemented. What has emerged over the last ten years is a cacophony of performance expectations that is confusing to all stakeholders. The time-honored tradition of state and local control in education cannot justify state standards so vastly disparate in their levels of difficulty. There is no reason to believe that the need for math or reading competence is any less in states like Wisconsin (whose standards are among the lowest we studied) than in South Carolina (whose standards are among the highest). Nor is it easy to explain why in many states, we see differences in standards that seem arbitrary across subjects. For example, Massachusetts adopted mathematics standards that would ensure all eighth-grade students are fully prepared for Algebra I, while adopting eighth-grade reading standards that do not ensure a minimum level of competence.

Figure 19 – State-reported mathematics proficiency rates for third grade, before and after calibration to the eighth-grade standards.



Standards have not remained consistent since NCLB's enactment, either. Some states have moved from highly challenging to less challenging standards, perhaps in response to NCLB requirements that 100 percent of students be proficient by 2014. A few states have raised the bar, setting higher standards and creating loftier expectations. These changes and inconstancies are part of a system of standards that fails to report student performance in a transparent manner and that makes tracking progress over time difficult. When states adopt new proficiency standards, stakeholders are routinely cautioned that prior achievement data are no longer relevant and that progress can be measured only using this new baseline.

Under the current system, standards are poorly calibrated across grades, which means that students who reach the proficiency standard in the early grades are often at risk of failing against the more challenging proficiency benchmarks of later grades. As we suggested earlier, this has created a misperception in some states that middle schools are performing worse than elementary schools, when in fact differences in proficiency rates are more often a product of differences in the relative difficulty of cut scores on state tests than of differences in performance.

Data from this study reinforce and echo findings from several other investigations that have found large disparities in the difficulty of state standards (National Center for Educational

Statistics 2007; Braun and Qian, 2005; Kingsbury et al. 2003; McGlaughlin and Bandiera de Mello 2003, 2002; McGlaughlin 1998a, 1998b). In particular, the findings of this study and those of the recent NCES study point toward the same general conclusions (see Appendix 8).

What would a better system look like? It would establish a single, national set of middle and high school performance expectations that would reflect the aspirations of most parents—including parents of historically disadvantaged minority groups—to have their children prepared to pursue post-secondary education. A recent New American Media poll of Latino, Asian, and African-American parents found that the vast majority expect their own children to graduate from a four-year university or attain a graduate degree (2006). The same group supported, by a very wide margin, a requirement that students pass exit examinations before receiving a high school diploma.

Such a standard could eventually be met by most students, although it would require rethinking the 100 percent proficiency requirement of NCLB. By establishing a single performance expectation that is aligned with college readiness, however, the system would more effectively communicate, especially to students and parents, whether a particular level of performance was sufficient to meet aspirations for the future. This would be a vast improvement over a system in which achieving a state's proficiency standard has little connection to preparedness for future education. It would also more effectively promote true educational equity and improve our national competitiveness.

An improved system would also exhibit consistency in the standards over time—a feature that would reflect constancy of purpose on the part of schools. One unintended consequence of NCLB has been the decision of some states—predominantly those that had established standards that seem to reflect college readiness—to lower their standards in order to meet NCLB requirements. In this context, constancy of purpose means not only maintaining a consistent level of difficulty on a test but also, more importantly, *maintaining a consistent purpose for the test itself.* In the past thirty years, educators have endured several waves of standards: first "minimum competency" standards, then "world-class" standards, then NCLB proficiency standards; and now there is the widespread call for standards reflecting some form of college readiness by the end of high school. One can understand if educators find these shifts confusing.

But regardless of what the final proficiency standards might be, the time has come for the proficiency standards to be final. Students, parents, educators, and other stakeholders have a right to know what the expectations are and how students are performing relative to them, and they need to know that the expectations are stable. This means that we cannot ease the standards if we discover that many students are not meeting performance goals. It may also mean that we may have to come up with a more sophisticated approach to accountability than the rather blunt instruments used by NCLB.

A strong accountability structure rests on three keystones. The first is high standards. The second is transparency, which ensures that the results produced by schools are properly documented, are made public, and are well-understood. The third keystone is a corrective system that *reliably* identifies schools performing poorly and implements whatever measures are needed to provide appropriate learning conditions for the students One of the major problems with NCLB lies with the third keystone. An accountability system that requires 100 percent of students to pass a test and puts all schools that fail to meet this standard on a path to closure is flawed because it does not reliably identify poor schools. Such a system is also politically unsustainable.

If state-level politicians are convinced that the rigor of their standards will force the closure of most of their schools, they may lower the standards and weaken the first keystone, or they may change the rules for adequate yearly progress, or engage in other coping mechanisms. These may delay sanctions, but they jeopardize the second keystone by making the results of the system less transparent.

Thus rather than strengthening accountability, the 100 percent requirement may have the opposite effect, both by making it difficult for states to sustain high standards for student performance, and by encouraging states to adopt rules for adequate yearly progress that make the system less transparent.

We believe that implementing a set of student proficiency standards that reflect the aspirations of parents is politically viable, and that reporting of performance relative to these standards can become more transparent. However, the 100 percent proficiency requirement and some of the other rules surrounding AYP must be changed. A more politically sustainable system is one that:

- Maintains standards for performance that reflect college readiness, in keeping with the hopes of parents and the needs of a post-industrial economy on a shrinking, flattening, and highly competitive planet

- Improves the transparency of the system by implementing more uniform rules governing AYP

- Creates accountability mechanisms to reward schools that produce high levels of performance and growth

- Supports schools that are making progress

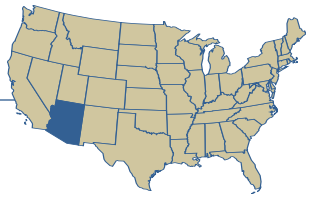
- Corrects or closes schools that clearly founder

Finally, an improved system of standards would be far more coherent than the one in place today. It would set expectations as high for reading as for mathematics. It would be designed to ensure that proficiency in the early grades is truly aligned with success in the upper grades. It would help parents know at any point in schooling whether their child's current performance and growth over time are on track to meet both their aspirations and the proficiency standards of the state. It would be structured so that schools get more reliable information about how students in the early grades are really performing relative to the school system's exit standards. In too many states, low proficiency standards in the early grades mask the true situation of youngsters who pass third-grade proficiency standards yet are not performing at a level that projects to success at later grades. Such children are truly at risk, yet invisible. A well-calibrated system of standards would address their situation and help schools allocate their resources to the areas of greatest student need.

The No Child Left Behind Act is worthy of praise for building a societal consensus around the premise that we should have high expectations for all of our children. While a certain amount of lip service was paid to this premise prior to NCLB, the bipartisan support for the act and the strong remedies associated with it communicate very clearly that the nation as a whole strongly supports educational equity.

What we have learned in five years, however, is that having expectations and sanctions is not sufficient. We also must have expectations that are consistent over time and place, coherent, and implemented in a manner that is politically sustainable. We have a national educational policy that is committed to "leave no child behind." The charge for Congress as it considers reauthorizing the act is to take the next large step toward fulfilling the expectation of students, parents, educators, and policymakers that our education system is prepared to help every student achieve his or her potential.

Arizona



Introduction

This study linked data from the 2002 and 2005 administrations of Arizona's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Arizona's definitions of "proficiency" in reading and mathematics are relatively consistent with the standards set by the other 25 states in this study. In other words, Arizona's tests are about average in terms of difficulty.

Yet the level of difficulty of Arizona's tests generally declined from 2002 to 2005—the No Child Left Behind era—quite significantly in some grades. This is not a surprise, as the Arizona State Board of Education adopted a new scale for both the reading and math tests for the 2004-05 academic school year, and publicly reported lowering the cut scores on those tests.

Not well known, however, is that the state's proficiency cut scores are now relatively lower for third-grade students than for eighth-grade pupils (taking into account the obvious differences in subject content and children's development). Plus, as is true for the majority of states studied, Arizona's cut scores for reading are lower than those for mathematics. Arizona policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating teacher and student performance across these domains.

What We Studied: Arizona's Instrument to Measure Standards (AIMS)

Arizona currently uses a spring assessment called the Arizona Instrument to Measure Standards – Dual Purposes Assessment (AIMS – DPA) as part of its state assessment program. This tests reading, writing, and mathematics in elementary and middle school students in grades 3 through 8. Students in grade 10 take the AIMS HS (High School) and may continue to take that test twice per year during grades 11 and 12 until they have met or exceeded the standards for proficiency in writing, reading, and mathematics. The current study analyzed reading and math results from a group of elementary and middle schools in which almost all students took both the state's assessment and MAP, using the spring 2002 and spring 2005 administrations of the two tests. (The methodology section of this report explains how performance on these two tests was compared.) These linked results were then used to estimate the scores on NWEA's scale that would be equivalent to the proficiency cut scores for each grade and subject on the Arizona State Assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered "proficient.")

Part 1: How Difficult are Arizona's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

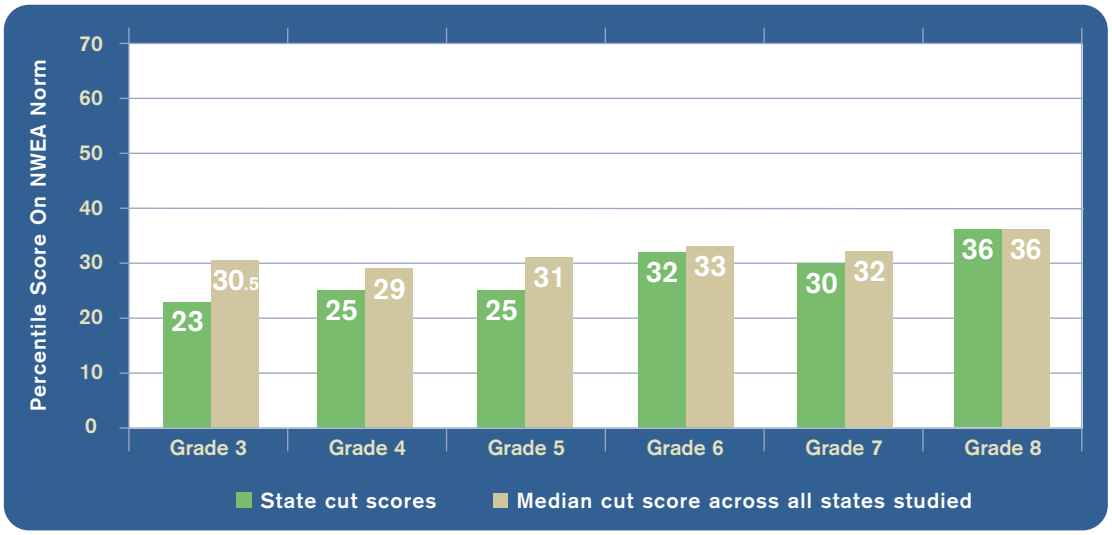
Applying that approach to this task, we evaluated the difficulty of Arizona's proficiency standards by estimating the proportion of students in NWEA's norm group who would perform above the Arizona standard on a test of equivalent difficulty. The following two figures show the difficulty of Arizona's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Arizona ranged between the 23rd and 36th percentiles for the norm group, with the eighth-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 28th and 42nd percentiles with eighth grade again being most challenging.

For most grade levels, Arizona's cut scores in both reading and mathematics are slightly below average in difficulty among the states studied. Exceptions include eighth-grade reading and sixth-grade math, which are at the median proficiency cut scores among the states examined.

Note, too, that Arizona's cut scores for reading are lower than those for mathematics. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Arizona students may be performing worse in reading and better in mathematics than is apparent by looking at the percentage of students passing state tests in those subjects.

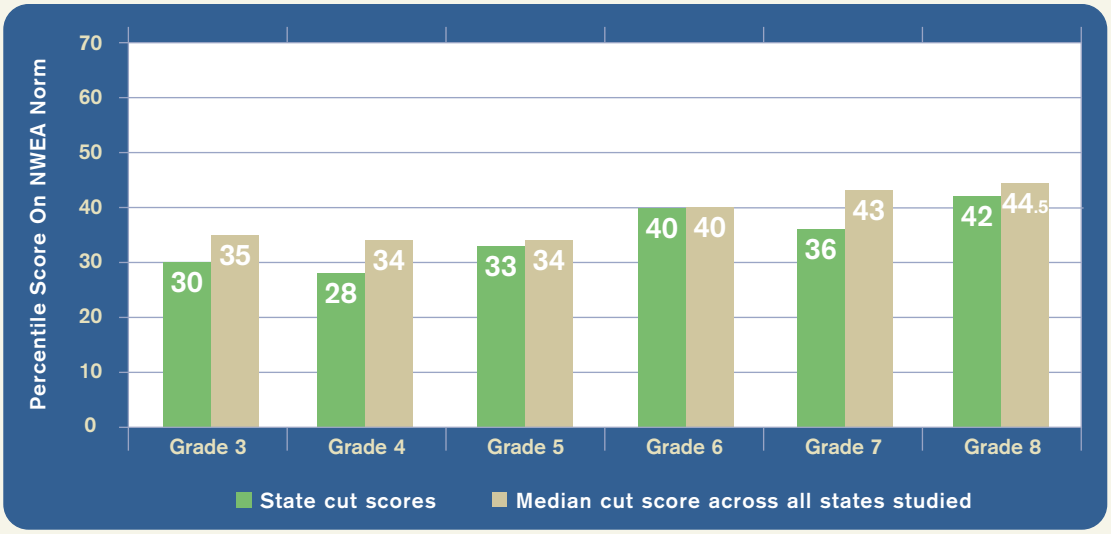
Another way of assessing difficulty is to evaluate how Arizona's proficiency cut scores rank relative to other states. Table 1 shows that the Arizona cut scores generally rank in the mid- or bottom third among the 26 states studied for this report. Arizona's third- and fifth-grade reading cut scores are particularly low, besting those of only seven other states in the study. On the other hand, Arizona ranks relatively high in eighth-grade math and reading and in third- and sixth-grade math.

Figure 1 – Arizona Reading Cut Scores in Relation to All 26 States Studied, 2005
(Expressed in 2005 MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the 2005 NWEA norm. These percentiles are compared with the median cut scores of other states reviewed in this study. Only in eighth grade does Arizona's cut score reach the median. Grades 3-7 scores are 1 to 7.5 percentile points below the median.

Figure 2 – Arizona Mathematics Cut Scores in Relation to All 26 States Studied, 2005 (Expressed in MAP Percentiles).



Note: Arizona's math test cut scores are shown as percentiles of the 2005 NWEA norm and compared with the median cut scores of other states reviewed in this study. Only in sixth grade does Arizona's cut score reach the median; in third grade, it lagged by 5 percentile points and in seventh grade by 7 points.

Table 1 – Arizona Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2005 or 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	19	17	19	14	18	9
Mathematics	14	19	16	12	18	12

Note: This table ranks Arizona's cut scores relative to the cut scores of the other 25 states in the study, where 1 is highest and 26 is lowest.

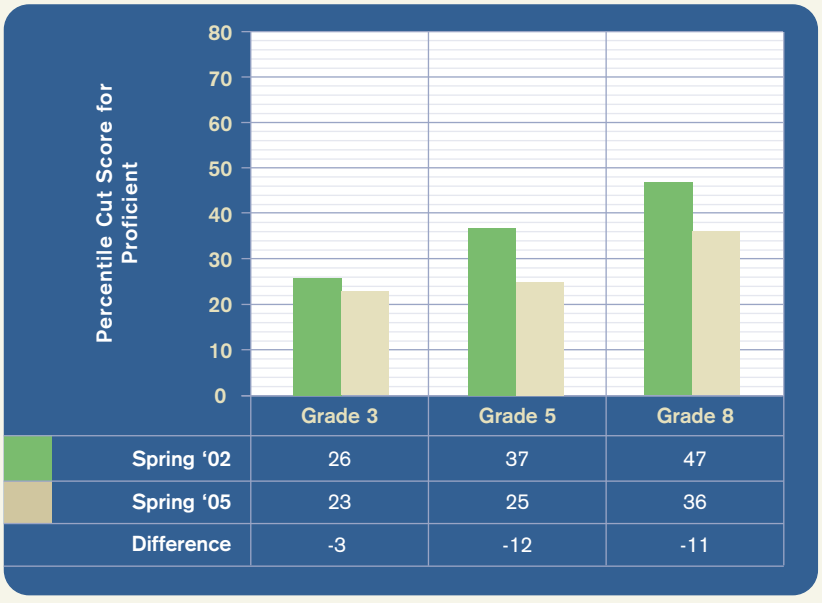
Part 2: Changes in Cut Scores over Time

In order to measure their consistency, Arizona's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2002 and 2005 school years. Cut score estimates for both years were available for grades 3, 5, and 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. This occurred in Arizona, in the 2004-05 academic year, when the State Board of Education adopted new scales and publicly lowered cut scores both for the reading and math tests.

Is it possible, then, to compare the proficiency scores between earlier administrations of Arizona's tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess. We can measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height to judge proficiency. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measure or scale used by the AIMS in 2002 and in 2005 can both be linked to the scale that was used to report MAP, which has remained consistent over time. Just as one can compare one meter to three feet and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the AIMS in 2002 and 2005 on the MAP scale and ascertain whether the test may have changed in difficulty—and whether those changes are consistent with what the state reported to the public.

Figure 3 – Estimated Change in Arizona's Proficiency Cut Scores in Reading, 2002-2005 (Expressed in MAP Percentiles).



	Grade 3	Grade 5	Grade 8
Spring '02	26	37	47
Spring '05	23	25	36
Difference	-3	-12	-11

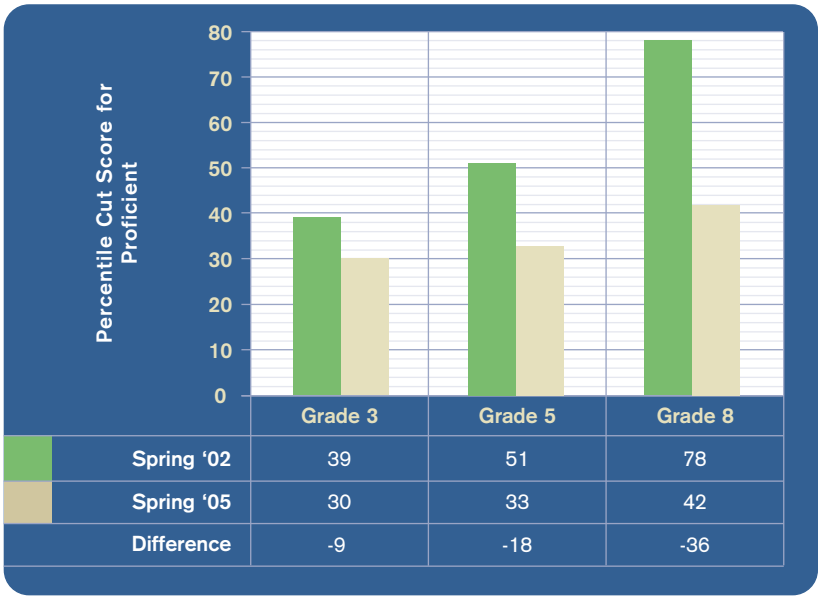
Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, fifth-grade students in 2002 had to score at the 37th percentile of the NWEA norm group in order to be considered proficient, while in 2005 fifth graders only had to score at the 25th percentile of the NWEA norm group to achieve proficiency. The change in grade 3 was within the margin of error (in other words, it is too small to be considered substantive).

Arizona's estimated **reading** cut scores decreased in grades 5 and 8 over this three-year period, though no substantive change was found in grade 3 (see Figure 3). Consequently, even though student performance on MAP did not change, one would expect the fifth- and eighth-grade reading proficiency rates in 2005 to be 12 percent and 11 percent higher than in 2002, respectively. (Arizona reported a 12-point gain for fifth graders and an 11-point gain for eighth graders over this period.)

Arizona's estimated **mathematics** cut scores indicate a dramatic decrease in proficiency cut scores in grades 3, 5, and 8 over this three-year period (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, the changes in grades 3, 5, and 8 would likely yield increased math proficiency rates of 9, 18, and 36 percent, respectively. Arizona reported a 15-point gain for third graders, a 25-point gain for fifth graders, and a 42-point gain for eighth graders over this period.)

Thus, one could fairly say that Arizona's third-grade reading test was about as difficult to pass in 2005 as in 2002, while the other tests were easier to pass for the other grades examined. As a result, some apparent improvements in the Arizona students' proficiency rates during this time may not be entirely a product of improved achievement.

Figure 4 – Estimated Differences in Arizona's Proficiency Cut Scores in Mathematics, 2002-2005 (Expressed in MAP Percentiles).



	Grade 3	Grade 5	Grade 8
Spring '02	39	51	78
Spring '05	30	33	42
Difference	-9	-18	-36

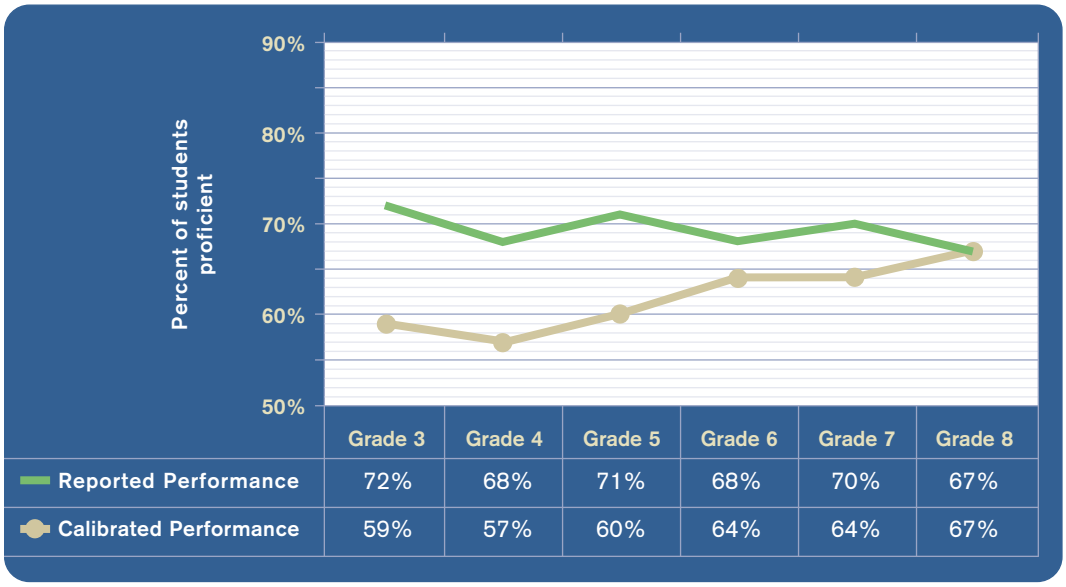
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, fifth-grade students in 2002 had to score at the 51st percentile of the NWEA norm group in order to be considered proficient, while in 2005 fifth graders only had to score at the 33rd percentile of the NWEA norm group to achieve proficiency.

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Arizona's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed that Arizona's upper grade cut scores in reading and mathematics in 2005 were more challenging than the cut scores in the lower grades. The two figures that follow show Arizona's reported performance on its state test in reading (Figure 5) and mathematics (Figure 6) compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade 8 standard. When differences in grade-to-grade difficulty of the cut scores are removed, student performance in mathematics is more consistent at all grades. This would lead to the conclusion that the higher rates of mathematics proficiency that the state has reported for elementary school students are somewhat misleading. It also becomes clear that actual reading performance is lower at the elementary level than in middle school—while the state's published passing rates appear to indicate relatively consistent performance from grades 3 to 8.

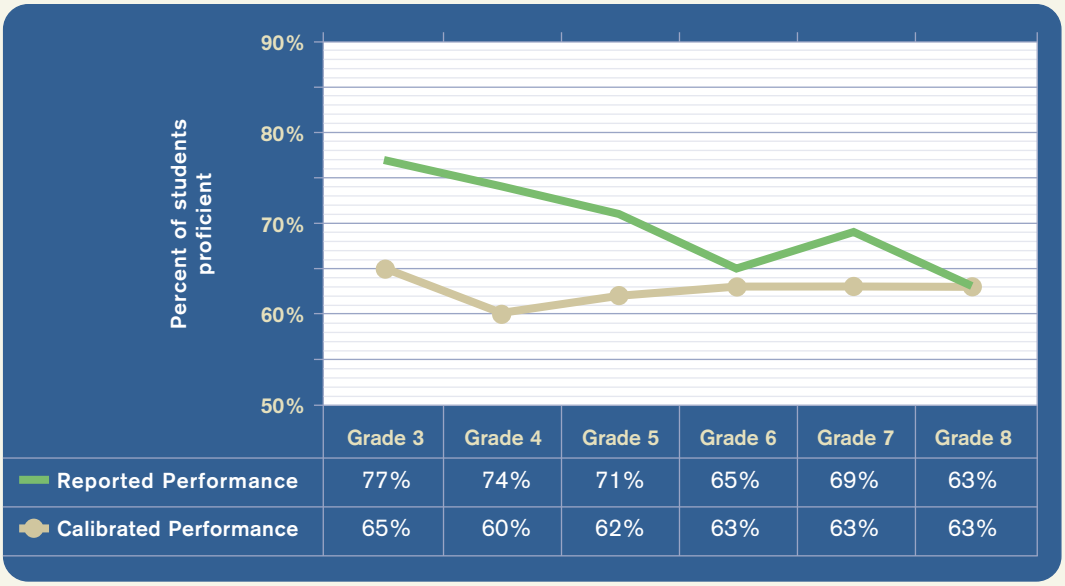
Figure 5 – Arizona Reading Performance as Reported and as Calibrated to the Grade 8 Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	72%	68%	71%	68%	70%	67%
Calibrated Performance	59%	57%	60%	64%	64%	67%

Note: This graphic shows, for example, that if Arizona's grade-3 reading standard were as difficult as its grade-8 standard, 59 percent of third graders would achieve the proficient level, rather than 72 percent, as reported by the state.

Figure 6 – Arizona Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



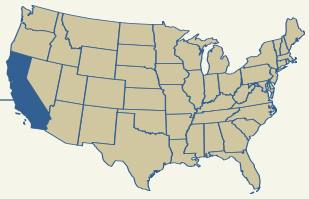
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	77%	74%	71%	65%	69%	63%
Calibrated Performance	65%	60%	62%	63%	63%	63%

Note: This graphic shows, for example, that if Arizona's grade-3 mathematics cut score were as difficult as its grade-8 standard, 65 percent of third graders would achieve the proficient level, rather than 77 percent, as was reported by the state.

Policy Implications

Arizona proficiency cut scores stand in the middle to bottom third of the pack when compared with the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales,* which also found Arizona's standards to be in the bottom half to the bottom third of the distribution of all states studied. Arizona's cut scores, which weren't particularly difficult in most grades in 2002, have over the past several years been adjusted—making them generally less challenging (and, in some grades, significantly less challenging). Arizona's expectations are not well calibrated across grades, particularly for mathematics. Students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. Arizona policymakers might consider adjusting their proficiency cut scores across grades so that parents and schools can be assured that young students scoring at the proficient level are truly prepared for success later in their educational careers.

California



Introduction

This study linked data from the 2003 and 2006 administrations of California's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that California's definitions of "proficiency" in reading and mathematics are relatively difficult compared with the standards set by the other 25 states in this study. In other words, it's harder to pass California's tests than those of most other states.

Yet, according to NWEA estimates, the difficulty level of California's tests declined between 2003 to 2006—the No Child Left Behind era. In a few grades, these declines were dramatic, calling into question some of the achievement gains previously reported by the state. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the California test not being matched by learning gains on the Northwest Evaluation Association test. Another interesting finding from this study is that California's mathematics proficiency cut scores are less stringent for third-grade students than they are for middle-school pupils (taking into account the obvious differences in subject content and children's development). California policymakers might consider adjusting their math cut scores to ensure equivalent difficulty at all grades so that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: California Standardized Testing and Reporting (STAR) Program

California currently uses a spring assessment called the California Standards Test (CST), which tests English/Language Arts and mathematics in grades 2 through 11. Students are also tested in science in grades 5, 8, and 10, and history in grades 8, 10, and 11. The current study analyzed reading and math results from a group of elementary and middle schools in which almost all students took both the state's assessment and MAP, using the spring 2003 and spring 2006 administrations of the two tests. (The methodology section of this report explains how performance on these two tests was compared.) These linked results were then used to estimate the scores on NWEA's scale that would be equivalent to the proficiency cut scores for each grade and subject on the CST (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.)

Part 1: How Difficult are California's Definitions of Proficiency in Reading and Math?

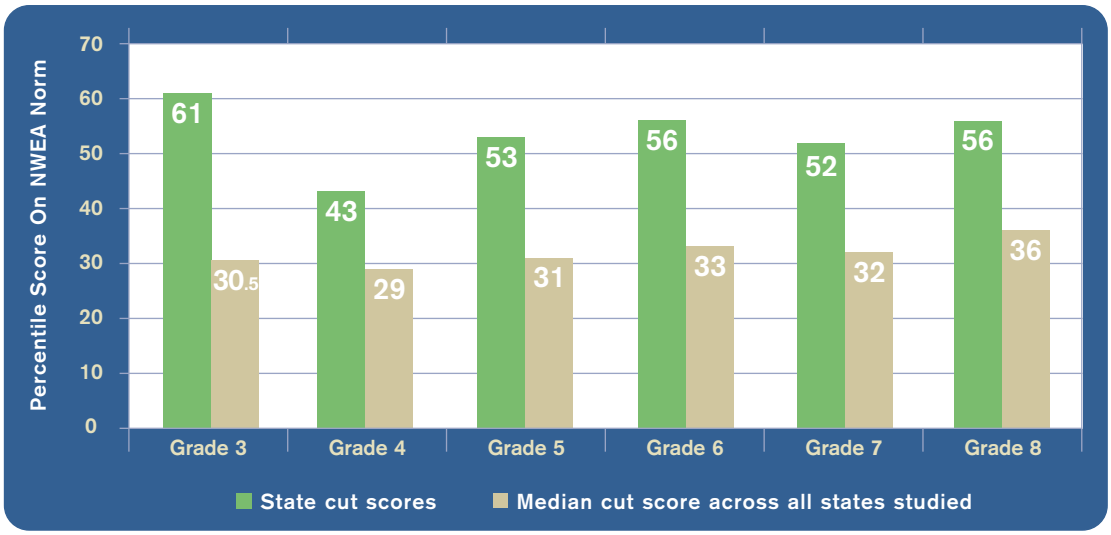
One way to assess the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of California's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the California standard on a test of equivalent difficulty. The following two figures show the difficulty of California's proficiency cut scores for **reading** (Figure 1) and **mathematics** (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for reading in California ranged between the 43rd and 61st percentiles for the norm group, with the third-grade cut score being most challenging. In mathematics, the proficiency cut scores ranged between 46th and 62nd percentiles, with sixth grade being most challenging. As is clear from Figures 1 and 2, California's cut scores in both reading and mathematics are consistently above average in difficulty among the states studied.

Note, too, that California's cut scores for reading tend to be slightly lower than the corresponding cut scores for mathematics at each grade, except for third grade. Thus, reported differences in achievement on the CST between reading and mathematics might be more a product of differences in cut scores than in actual student achievement. In other words, California students may be performing worse in reading or better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

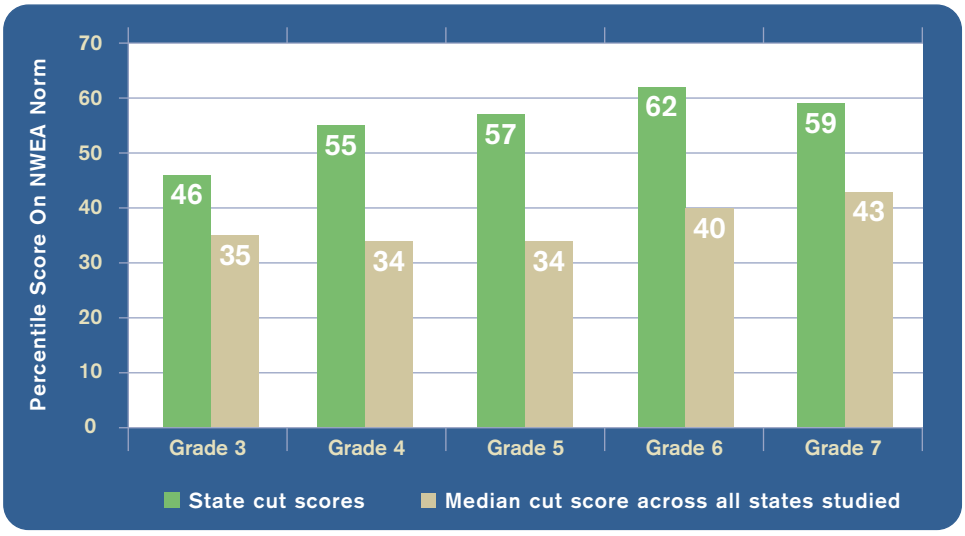
Another way of assessing difficulty is to evaluate how California's proficiency cut scores rank relative to other states. Table 1 shows that the California cut scores generally rank near the top of the 26 states studied for this report. Its reading cut score in grade 3 ranks first across all states within the current study.

Figure 1 – California Reading Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in 2005 MAP Percentiles)



Note: This figure shows California's 2006 reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of all 26 states reviewed in this study. California's cut scores are consistently 14 to 30.5 percentiles above the median in grades 3-8.

Figure 2 – California Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in 2005 MAP Percentiles)



Note: California's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. California's cut scores in grades 3-6 are consistently 11 to 23 percentiles above the median.

Table 1 – Ranking of 2006 California Reading and Mathematics Cut Scores for Proficient Performance in Relation to All States Studied

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	1	3	2	2	2	2
Mathematics	4	3	3	3	4	Not available

Note: This table ranks California's cut scores relative to the cut scores of the other 25 states in the study. For third-grade reading, California ranks 1 out of 26, meaning that California's cut scores were the highest of the states studied.

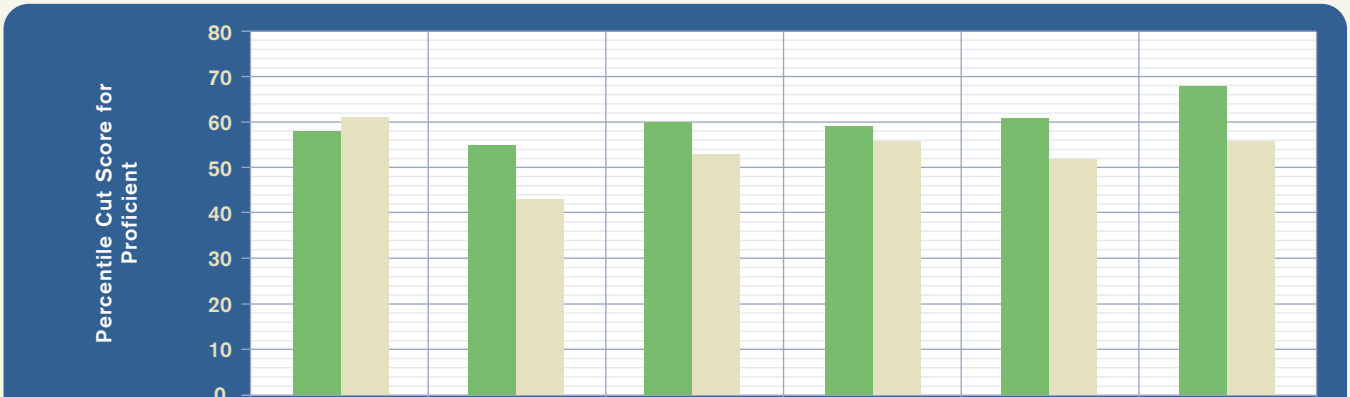
Part 2: Changes in Cut Scores over Time

In order to measure their consistency over time, California's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2003 and 2006 school years. Cut score estimates for the three-year duration are available for reading in grades 3 through 8, and grades 3 through 7 for mathematics.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the tests used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Plus, unintentional drift can occur even in states, such as California, that maintained their proficiency levels.

Is it possible, then, to compare the proficiency scores between earlier administrations of California tests with today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measure or scale used by the CST in 2003 and in 2006 can be linked to the scale used for MAP, which has remained consistent over time. Just as one can compare three feet to a meter and know that a one meter jump is slightly more difficult than a three foot jump, one can estimate the cut score needed to pass the CST in 2003 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty.

Figure 3 – Estimated Differences in California's Proficiency Cut Scores in Reading, 2003-2006 (Expressed in MAP Percentiles).



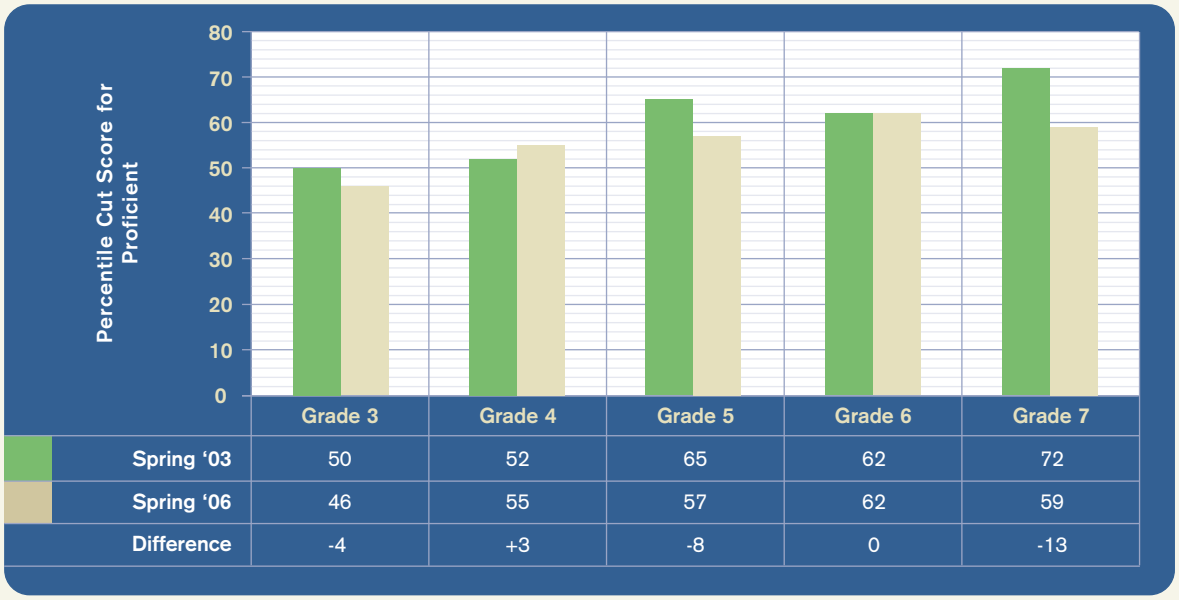
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Spring '03	58	55	60	59	61	68
Spring '06	61	43	53	56	52	56
Difference	+3	-12	-7	-3	-9	-12

Note: This graphic shows how the degree of difficulty in achieving proficiency in reading has changed. For example, eighth-grade students in 2003 had to score at the 68th percentile of the NWEA norm group in order to be considered proficient, while in 2006 eighth graders only had to score at the 56th percentile to achieve proficiency. The changes in grades 3, 5, and 6 were within the margin of error (in other words, too small to be considered substantive).

Despite the fact (see Figures 1 and 2) that California's 2006 cut scores were among the most challenging in the country, the state's estimated **reading** cut scores decreased substantially in fourth, seventh, and eighth grades over this three-year period (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the fourth, seventh, and eighth grade reading proficiency rates in 2006 to be 12 percent, 9 percent, and 12 percent higher than in 2003, respectively. California reported a 10 point gain for fourth graders, a 7 point gain for seventh graders, and a 11 point gain for eighth graders over this period.

California's estimated **mathematics** results indicate a decrease in proficiency cut scores in grades 5 and 7 over this three-year period (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, the changes in grades 5 and 7 would likely yield increased pupil proficiency rates of 12 percent and 13 percent, respectively. (California reported a 13 point gain for fifth graders and an 11 point gain for seventh graders over this period.) Thus, one could fairly say that California's seventh-grade tests in both reading and mathematics were easier to pass in 2006 than in 2003, while third and sixth grade tests were about the same. As a result, improvements in state-reported proficiency rates for grades whose tests became easier may not be entirely a product of improved achievement.

Figure 4 – Estimated Differences in California's Proficiency Cut Scores in Mathematics, 2003-2006 (Expressed in MAP Percentiles).



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Spring '03	50	52	65	62	72
Spring '06	46	55	57	62	59
Difference	-4	+3	-8	0	-13

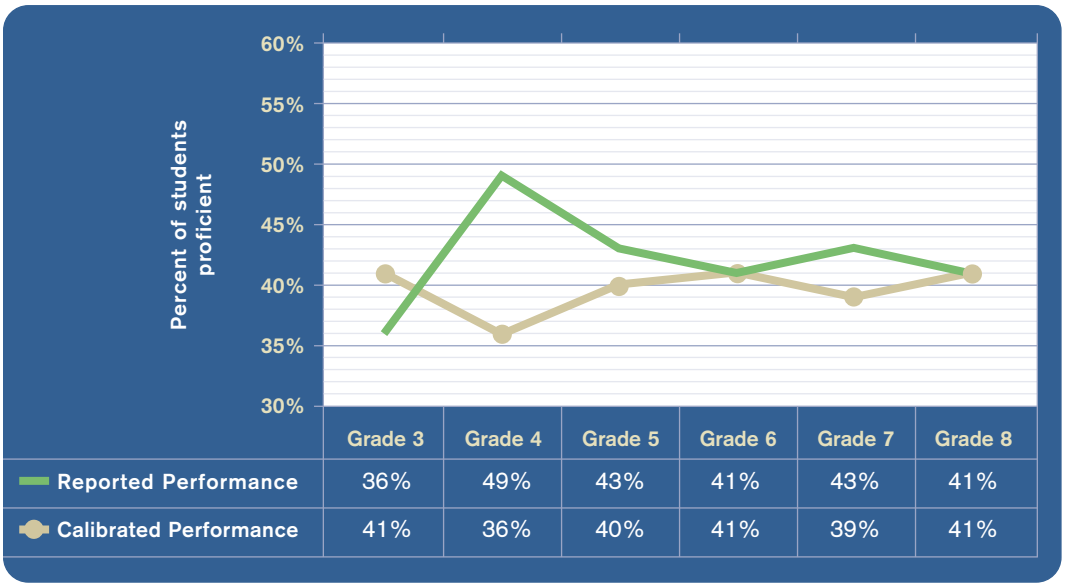
Note: This graphic shows how the degree of difficulty in achieving proficiency in math has changed. For example, seventh-grade students in 2003 had to score at the 72nd percentile of the NWEA norm group in order to be considered proficient, while by 2006 seventh graders had only to score at the 59th percentile to achieve proficiency. The changes in grades 3, 4, and 6 were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining California's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed that California's third-grade reading cut score in 2006 was more challenging than reading cut scores in higher grades, but that the third-grade mathematics cut score was lower than in subsequent grades. The two figures that follow show California's reported performance on its state test in reading (Figure 5) and mathematics (Figure 6) compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-eight standard. When differences in grade-to-grade difficulty of the cut scores are removed, student performance in mathematics is more consistent at all grades.

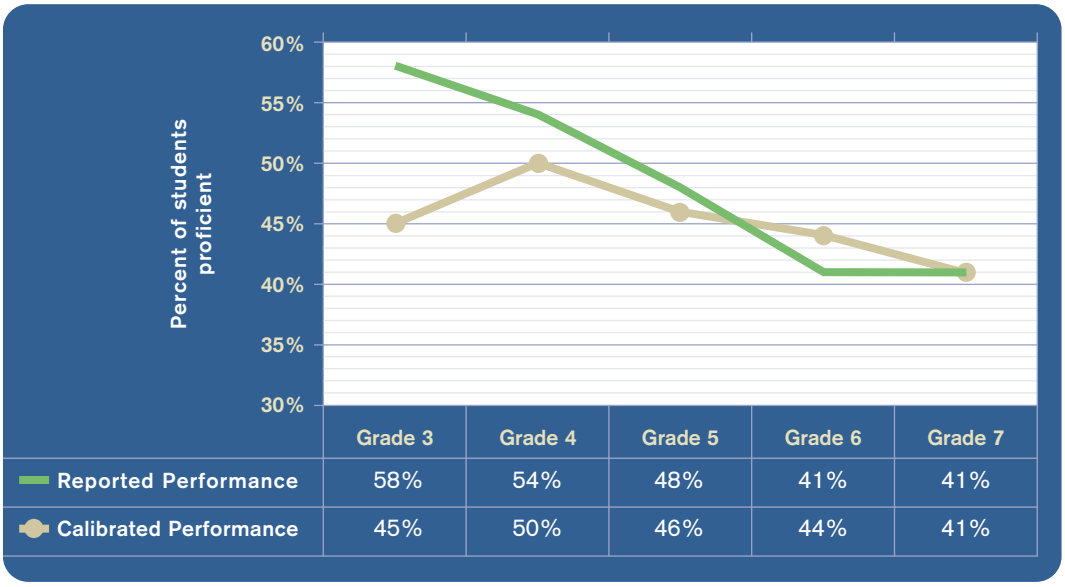
Figure 5 – California Reading Performance as Reported and as Calibrated to the Grade 8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	36%	49%	43%	41%	43%	41%
Calibrated Performance	41%	36%	40%	41%	39%	41%

Note: This graphic means that, for example, if California's third-grade reading standard was set at the same level of difficulty as its eighth-grade reading standard, 41 percent of third graders would achieve the proficient level, rather than 36 percent, as reported by the state.

Figure 6 – California Mathematics Performance as Reported and as Calibrated to the Grade 8 Standard, 2006



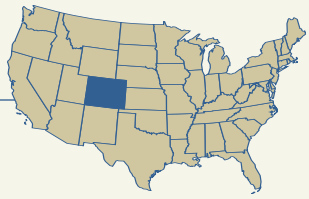
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Reported Performance	58%	54%	48%	41%	41%
Calibrated Performance	45%	50%	46%	44%	41%

Note: This graphic means that, for example, if California's third-grade mathematics standard was as rigorous as its eighth-grade standard, 44 percent of third graders would achieve the proficient level, rather than 57 percent, as reported by the state.

Policy Implications

California's proficiency cut scores are very challenging when compared with the other 25 states in this study, ranking near the top. This finding is relatively consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales,* which also found California's cut scores to be near the top of the distribution of all states studied. Yet California's cut scores have changed over the past several years—making them generally less challenging, in some cases dramatically so, though not in all grades. As a result, California's expectations are not smoothly calibrated across grades; students who are proficient in third-grade math, for example, are not necessarily on track to be proficient in the eighth grade. California policymakers might consider adjusting their mathematics cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Colorado



Introduction

This study linked data from the 2002 and 2005 administrations of Colorado's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that, for purposes of complying with the federal No Child Left Behind Act (NCLB), Colorado's definitions of "proficiency" in reading and mathematics are much less difficult than the standards set by most of the other 25 states in this study. In other words, it's easier to pass Colorado's tests than those of almost all other states.

Moreover, the difficulty of Colorado's tests decreased somewhat from 2002 to 2005—the NCLB era—although not for all grades. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the Colorado test not being matched by learning gains on the Northwest Evaluation Association test. One finding of this study is that Colorado's cut scores are now relatively less difficult at the lower grades than at the higher ones (taking into account the obvious differences in subject content and children's development). Colorado policymakers might consider raising their standards in the earlier grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

In this study, we used the proficiency cut scores that Colorado employs for purposes of NCLB to make comparisons. It's well known that Colorado opted to use the state's *partially proficient* level of academic performance as *proficient* for NCLB purposes. Hence we follow that practice here and subsequent references to "proficient" or "proficiency" in Colorado should be understood accordingly.

What We Studied: Colorado Student Assessment Program (CSAP)

Colorado currently uses an assessment called the Colorado Student Assessment Program (CSAP) which tests reading, writing, and math in grades 3-10 and science in grade 8. The same sets of tests were used in spring 2002 in which reading and writing were administered in grades 3-10, while math was administered in grades 5-10, and science was administered in grade 8. The current study linked data from spring 2002 and spring 2005 CSAP administrations to MAP, which was also administered in the 2002 and 2005 school years and has an unchanging scale.

To estimate the difficulty of Colorado's proficiency cut scores, we linked data from Colorado's reading and math tests from a group of elementary and middle schools to the NWEA assessment. (A "proficiency cut score" is the test score that a student must achieve in order to be considered proficient.) This was done by analyzing a group of schools in which almost all students had taken both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Colorado's Definitions of Proficiency in Reading and Math?

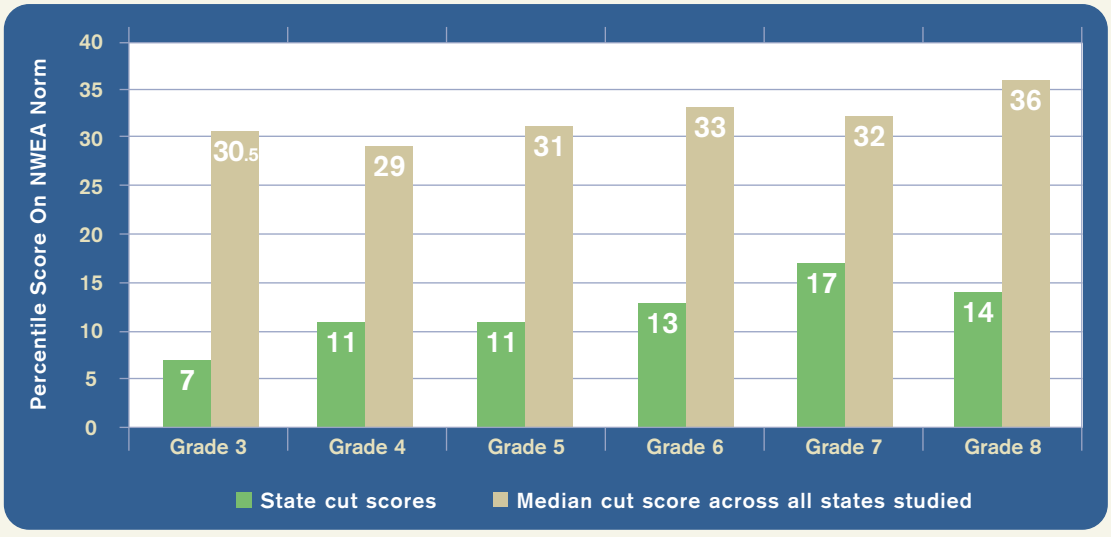
One way to assess the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. How do we know that solving differential equations is more difficult than adding fractions? Because if you ask a group of tenth graders to do both tasks, far more will be able to add fractions than will be able to solve differential equations.

Applying that approach to this task, we evaluated the difficulty of Colorado's NCLB proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Colorado cut score on a test of equivalent difficulty. The following two figures show the difficulty of Colorado's proficiency cut scores for **reading** (Figure 1) and **mathematics** (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The NCLB proficiency cut scores for reading in Colorado ranged between the 7th and 17th percentiles for the norm group, with the seventh grade being most challenging. In mathematics, the NCLB proficiency cut scores ranged between the 6th and 25th percentiles for the norm group with the eighth grade being most challenging.

Colorado's NCLB cut scores in both reading and mathematics are well below average in difficulty among the states studied. Note, too, that in middle school, Colorado's cut scores for reading are lower than those for mathematics. Thus, reported differences in achievement on the CSAP between reading and mathematics might be more a product of differences in cut scores than in actual student achievement. In other words, Colorado students might be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

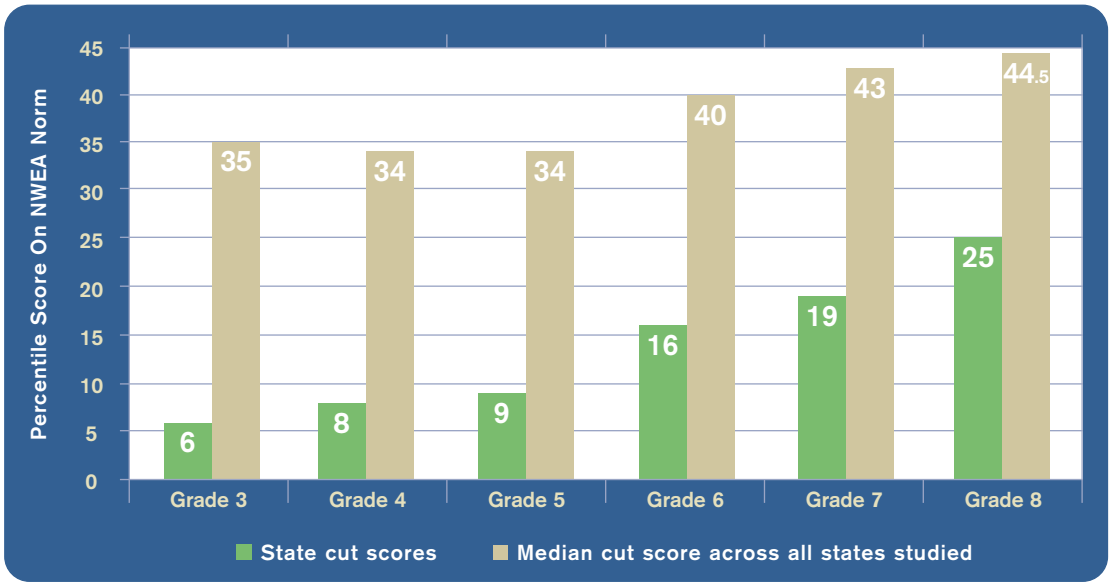
Another way of assessing difficulty is to evaluate how Colorado's NCLB proficiency cut scores rank relative to other states. Table 1 shows that the Colorado cut scores generally rank among the lowest of the 26 states studied for this report. In third and fifth grade reading, Colorado's cut scores rank; the state is second-to-last in fourth, sixth, and seventh grade reading and fifth grade mathematics.

Figure 1 – Estimate of Colorado Reading Cut Scores in Relation to the 25 Other States Studied, 2006 (Expressed in MAP Percentile Ranks)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NEWA norm. These percentiles are compared with the media cut scores of all 26 states reviewed in this study. Colorado's cut scores are consistently 15 to 23.5 percentile points below the median in grades 3 to 8.

Figure 2 – Colorado Mathematics Cut Scores in Relation to the 25 Other States Studied, 2006 (as Expressed in MAP Percentile Ranks)



Note: Colorado's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. Colorado's cut scores are 29 to 19.5 percentiles below the median across grades 3-8.

Table 1 – Colorado Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	26	25	26	25	25	23
Mathematics	24	24	25	24	23	19

Note: This table ranks Colorado's cut scores relative to the cut scores of the other 25 states in the study. In third-grade math, Colorado ranks 24 out of 26, meaning that 23 states' cut scores were higher, while only two were lower. Colorado either places last or second-to-last in half the categories.

Part 2: Changes in Cut Scores over Time

In order to measure their consistency over time, Colorado's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2002 and 2005 school years. Cut score estimates for both years were available for grades 3-8 for reading, and grades 5-8 for mathematics.

States may periodically re-adjust the cut scores they use to define proficiency in reading and mathematics, or update the tests used to evaluate student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed.

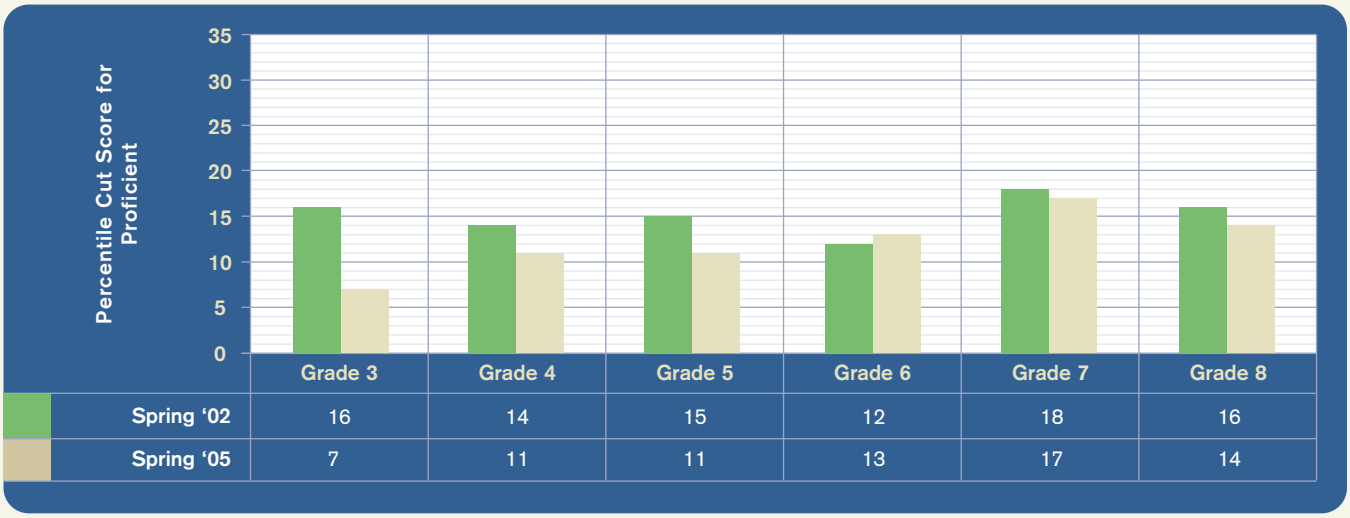
Is it possible, then, to compare the proficiency scores between the earlier era of Colorado's tests and today's? Yes. Assume once again that we're judging a group of fourth graders on their high-jump ability and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at 1 meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear 1 meter than 3 feet, because we know the relationship between the measures. The same principle applies here. CSAP in 2002 and in 2005 can both be linked to the MAP, which has remained consistent over time. Just as one can convert three feet to a meter [see comments in CA write up] and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the CSAP in 2002 and 2005 on the MAP scale and ascertain whether the test may have changed in difficulty.

Colorado's **reading** results indicate a decline in estimated proficiency cut scores in grades three, four, and five over this three-year period (see Figure 3). Consequently, one would expect the third grade students' reading proficiency rates in 2005 to be 9 percent higher than in 2002, even if actual pupil student performance remained the same. One would expect similar increases in the reading proficiency rates for fourth and fifth grades of 3 and 4 percent, respectively, if actual student performance remained the same.

Colorado's **mathematics** results indicate a decrease in estimated proficiency cut scores in grades 5, 7, and 8 (Figure 4). These changes would likely yield increased math proficiency rates in these grades of 4, 5, and 6 percent, respectively, even if pupil performance remained the same.

Thus, one could fairly say that Colorado's fifth grade tests in both reading and mathematics were easier to pass in 2005 than in 2002. Similarly, the reading tests for third and fourth graders were easier, as were the mathematics tests for seventh and eighth graders. As a result, some apparent improvements in Colorado students' proficiency rates during this period may not be entirely a product of improved achievement.

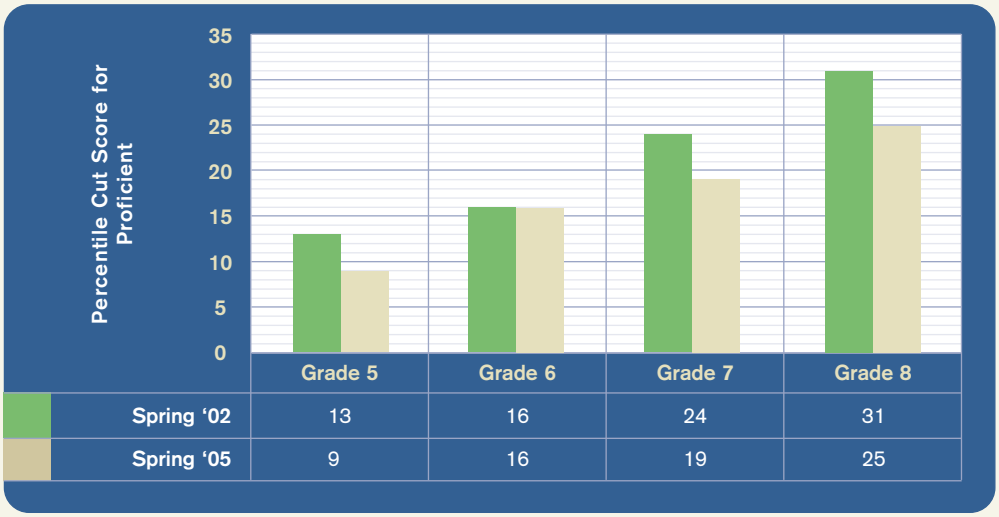
Figure 3 – Estimated Differences in Colorado's Proficiency Cut Scores in Reading, 2002-2005 (Expressed in MAP Percentile Ranks).



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Spring '02	16	14	15	12	18	16
Spring '05	7	11	11	13	17	14

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third grade students in 2002 had to score at the 16th percentile in order to be considered proficient, while in 2005 third graders had only to score at the 7th percentile.

Figure 4 – Estimated Differences in Colorado's Proficiency Cut Scores in Mathematics, 2002-2005 (Expressed in MAP Percentile Ranks).



	Grade 5	Grade 6	Grade 7	Grade 8
Spring '02	13	16	24	31
Spring '05	9	16	19	25

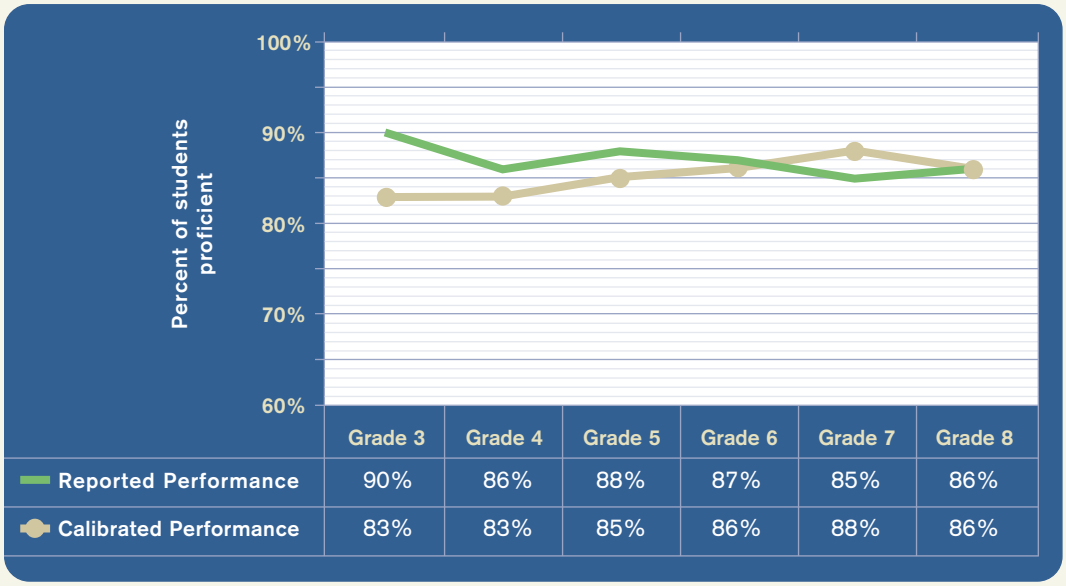
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, fifth grade students in 2002 had to score at the 13th percentile in order to be considered proficient, while by 2005 fifth graders only had to score at the 9th percentile.

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Colorado's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed that Colorado's upper-grade cut scores in reading and mathematics in 2005 were more challenging than in the lower grades. The two figures that follow show Colorado's reported performance on its state test in reading (Figure 5) and mathematics (Figure 6) compared with the rates of proficiency that would be achieved if the cut scores were calibrated to grade 8. When differences in grade-to-grade difficulty of the cut scores are removed, student performance is more consistent at all grades, particularly in mathematics. This would lead to the conclusion that the higher rates of mathematics proficiency that the state has reported for younger students are somewhat misleading.

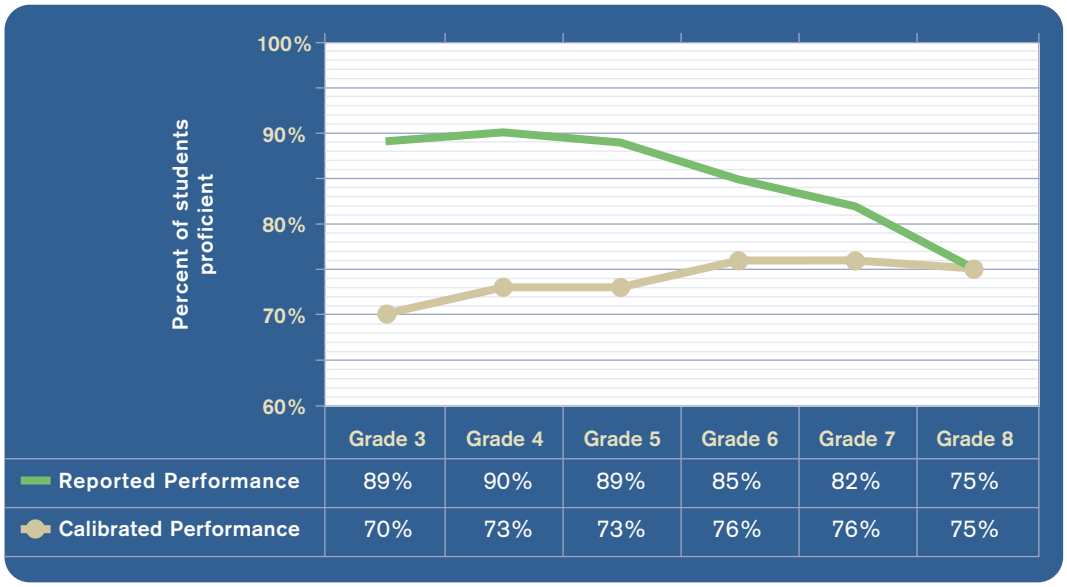
Figure 5 – Colorado Reading Performance Relative to a Calibrated Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	90%	86%	88%	87%	85%	86%
Calibrated Performance	83%	83%	85%	86%	88%	86%

Note: This graphic shows, for example, that if Colorado's grade 3 reading standard were as difficult as its grade 8 standard, 83 percent of third graders would achieve the proficient level, rather than 90 percent, as was reported by the state.

Figure 6 – Colorado Mathematics Performance Relative to a Calibrated Standard, 2005



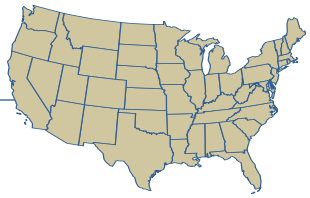
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	89%	90%	89%	85%	82%	75%
Calibrated Performance	70%	73%	73%	76%	76%	75%

Note: This graphic shows, for example, that if Colorado's grade 3 mathematics standard were set at the same level of difficulty as its grade 8 standard, 70 percent of third graders would achieve the proficient level, rather than 89 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what constitutes student proficiency in reading and mathematics for NCLB purposes, Colorado aimed low, at least compared to the other 25 states in this study. (This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Colorado's standards to be toward the bottom of the distribution of all states studied.) Colorado's low cut scores have declined even further in recent years in several grades. As a result, Colorado's expectations are not calibrated across all grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. In addition to better calibrating the state's cut scores, Colorado policymakers might consider raising those scores across the board so that parents and educators can be assured that scoring at the NCLB proficient level means that students are truly prepared for success later in their educational careers.

Delaware



Introduction

This study linked data from the 2006 administration of Delaware's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Delaware's definitions of proficiency in reading and mathematics generally ranked below average compared with the standards set by the 25 other states in this study.

Moreover, Delaware's proficiency cut scores in math are relatively lower in early grades than in later grades (taking into account the obvious differences in subject content and children's development). Therefore, reported results may overestimate the number of elementary students on track to be proficient in math by the eighth grade. Delaware policymakers might consider adjusting their math standards to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Delaware Student Testing Program (DSTP)

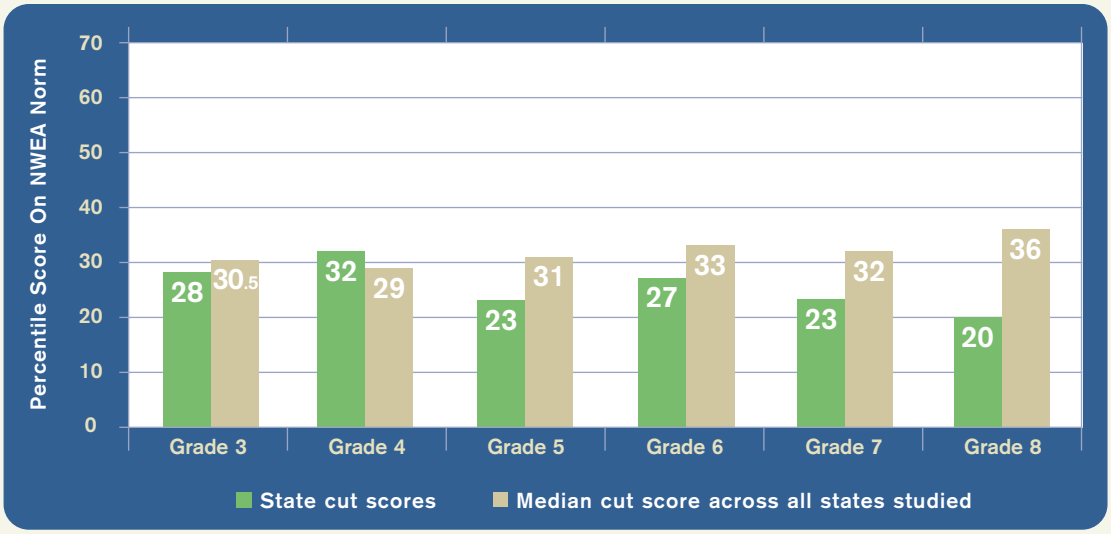
Delaware currently uses an assessment called the Delaware Student Testing Program (DSTP), which tests reading, writing, and mathematics in grades 2-10. The current study analyzed reading and math results from a group of elementary and middle schools in which almost all students had taken both the state assessment and MAP, using the spring 2006 administrations of the two tests. (The methodology section of this report explains how performance on these two tests was compared.) These linked results were then used to estimate the scores on NWEA's scale that would be equivalent to the proficiency cut scores for each grade and subject on the Delaware State Assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.)

Part 1: How Difficult are Delaware's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to leap? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? We know because only one (or perhaps none) of those same 100 individuals would successfully meet that level of challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

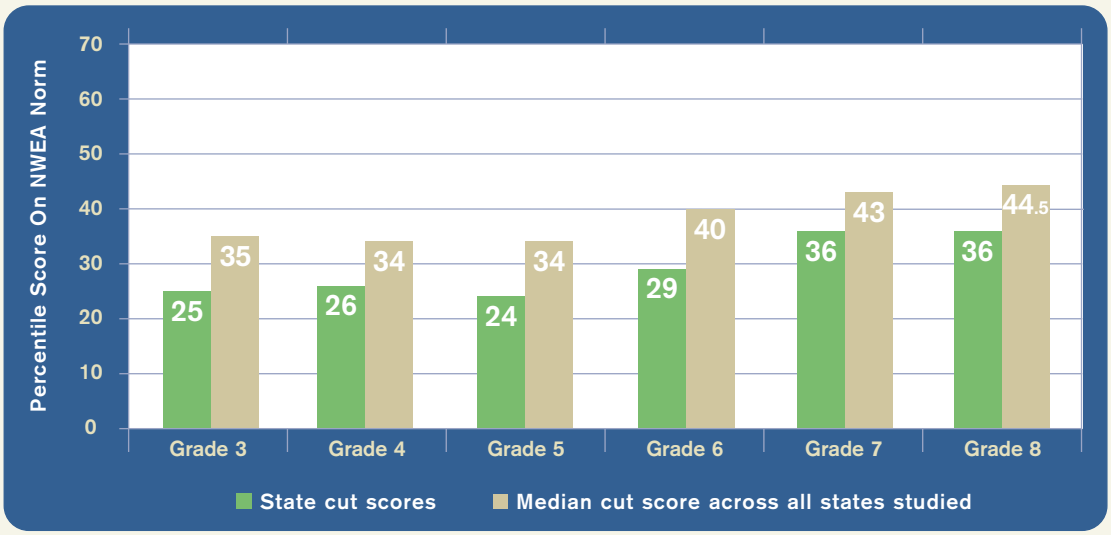
Applying the concept to this task, we evaluated the difficulty of the Delaware proficiency cut scores by estimating the proportion of students in NWEA's norm group that would perform above the Delaware standard on a test of equivalent difficulty. The following two figures show the difficulty of Delaware's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Delaware ranged between the 20th and 32nd percentiles for the norm group, with the fourth-grade standard being most challenging. In **mathematics**, the proficiency cut scores ranged between the 24th and 36th percentiles with seventh and eighth grade being most challenging.

Figure 1 – Delaware Reading Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of other states reviewed in this study. Only in fourth grade does Delaware surpass the median; by eighth grade, its reading cut score is 16 percentiles below the median.

Figure 2 – Delaware Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles).



Note: Delaware's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. The proficiency cut scores are consistently 7 to 11 percentiles below the median.

Table 1 – 2006 Delaware Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	14	10	20	18	22	22
Mathematics	20	21	20	20	18	16

Note: This table ranks Delaware's cut scores relative to the cut scores of the other 26 states in the study where 1 is the highest rank and 26 is the lowest.

Delaware's cut scores in reading and math are below average in difficulty for most grades, compared with other states in the study. The reading proficiency cut scores are also lower than those for mathematics. (This was the case for the majority of states studied.) Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Delaware students may be performing worse in reading and/or better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

Another way of assessing difficulty is to evaluate how Delaware's proficiency cut scores rank relative to other states. Table 1 shows that the Delaware proficiency cut scores generally rank in the middle to lower third in difficulty among the 26 states studied for this report; its cut scores are especially low for seventh-and eighth-grade reading.

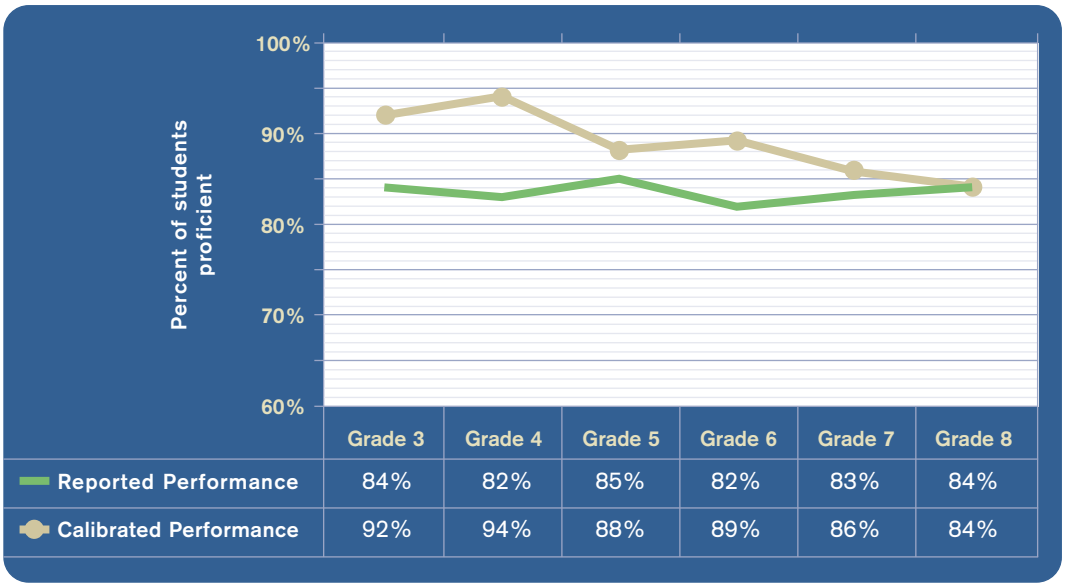
Part 2: Calibration across Grades*

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Delaware's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 above showed that Delaware's reading and mathematics proficiency cut scores in 2006 differed across grades in terms of their relative difficulty. The two figures that follow show Delaware's reported performance on its state test in reading (Figure 3) and mathematics (Figure 4), compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When the differences in grade-to-grade difficulty of the cut scores are removed, student performance is more consistent at all grades, at least in math.

*Delaware was one of seven states in this study for which cut score estimates could be reported for only a single year (2006). Eighth-grade cut score estimates for math and reading for the 2005 year were computed for Delaware, but it was determined that this single-grade estimate would be insufficient to draw overall conclusions about changes over time for the state. Consequently, changes over time are not included in Delaware's state report.

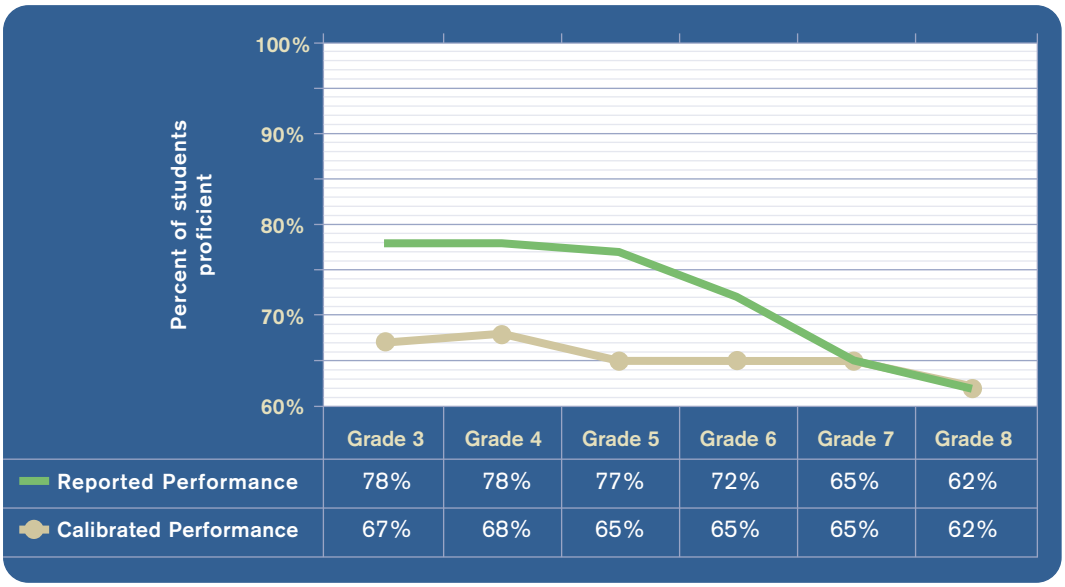
Figure 3 – Delaware Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	84%	82%	85%	82%	83%	84%
Calibrated Performance	92%	94%	88%	89%	86%	84%

Note: This graphic shows, for example, that, if Delaware's grade-3 reading standard were set at the same level of difficulty as its grade-8 standard, 92 percent of third graders would achieve the proficient level, rather than 84 percent, as reported by the state.

Figure 4 – Delaware Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



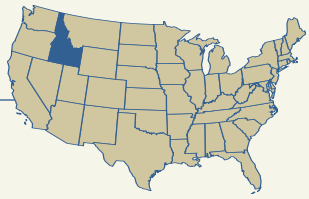
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	78%	78%	77%	72%	65%	62%
Calibrated Performance	67%	68%	65%	65%	65%	62%

Note: This graphic shows, for example, that if Delaware's grade-3 mathematics standard were as difficult as its grade-8 standard, 67 percent of third graders would achieve the proficient level, rather than 78 percent, as was reported by the state.

Policy Implications

Delaware's proficiency cut scores are in the middle to lower end of the pack when compared with the other 25 states in this study. (This finding is relatively consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Delaware's reading standards to be in the bottom half to the bottom third of the distribution of states studied and its math standards to be about in the middle.) In addition, Delaware's expectations in math are not smoothly calibrated across grades; students who are proficient in third-grade math are not necessarily on track to be proficient by the eighth grade. Delaware policymakers might consider adjusting their math cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating teacher and student performance across these domains.

Idaho



Introduction

This study used data from the 2002 and 2006 administrations of Idaho's state reading and math tests. We found that, compared with the other 25 states in this study, Idaho's definition of "proficiency" in reading and mathematics is relatively consistent with the cut scores set by other states. In other words, Idaho's tests are about average in terms of difficulty. However, Idaho's cut scores for third-grade mathematics are less difficult than they are for eighth-grade students, meaning that the state might be overstating the number of younger students who are actually on track academically. Idaho policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Idaho Standards Achievement Tests (ISAT)

Idaho currently uses the Idaho Standards Achievement Tests (ISAT), which test students in grades 2 through 10 in reading, mathematics, and language usage. Science is also tested in grades 5, 7, and 10. The version of ISAT administered during the study period was derived from NWEA's Measures of Academic Progress (MAP) and constructed specifically for use with students in Idaho. The current study shows how proficiency levels in Idaho, as determined by cut scores on the ISAT/MAP, compare with the cut scores in use in other states. Because Idaho used NWEA's scale for its state assessment, Idaho's proficiency cut scores could be compared directly to those of other states without need to convert cut scores.

Part 1: How Difficult are Idaho's Definitions of Proficiency in Reading and Math?

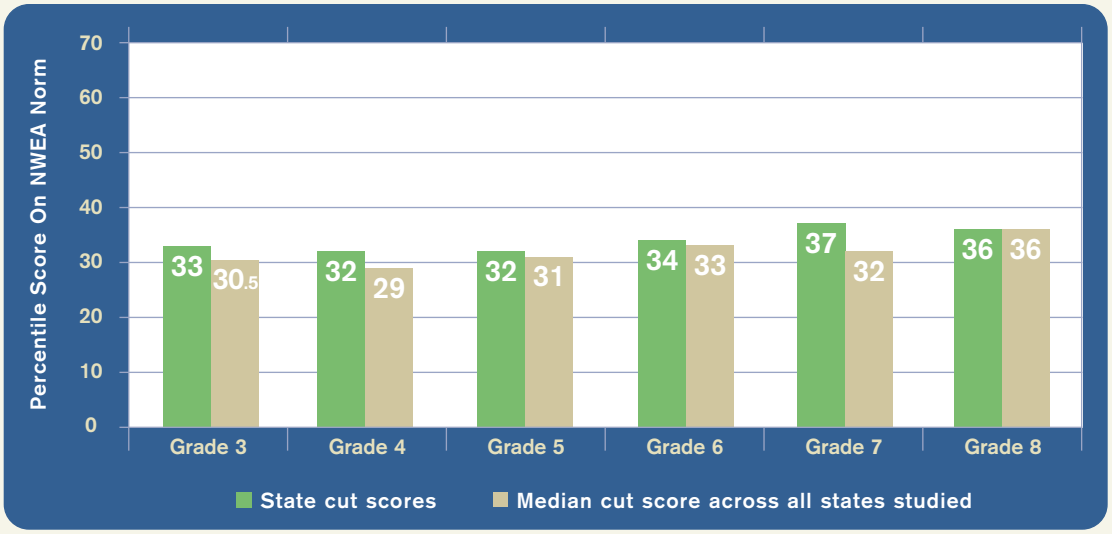
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

We evaluated the difficulty of Idaho's proficiency cut scores by estimating the proportion of students in NWEA's multi-state norm group who would perform above the Idaho cut score on a test of equivalent difficulty. The following two figures show the difficulty of Idaho's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Idaho range between the 32nd and 37th percentiles with respect to the NWEA norm group, with the seventh grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 30th and 47th percentiles, with the eighth grade being most challenging.

Idaho's cut scores for reading and mathematics tend to fall at about the median level of difficulty among the 26 states studied. Note, too, that the difficulty of Idaho's reading cut scores is lower than the corresponding mathematics cut scores except in third grade. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut score difficulty than in actual student achievement. In other words, Idaho students may be performing worse in reading and better in mathematics than is apparent by looking at the percentage of students passing state tests in those subjects.

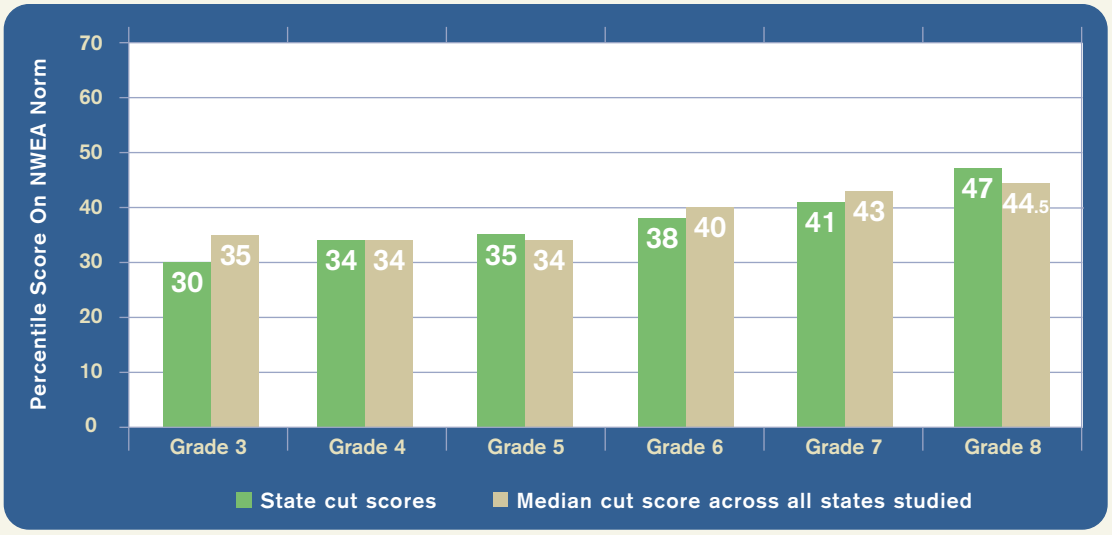
Another way of assessing difficulty is to evaluate how Idaho's proficiency cut scores rank relative to other states. Table 1 shows that the Idaho cut scores generally rank in the middle third in difficulty among the 26 states studied for this report.

Figure 1 – Idaho Reading Cut Scores in Relation to All 26 States Studied, 2006
(expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. Idaho's percentiles are compared with the median cut scores of all 26 states reviewed in this study. Idaho's cut scores are consistently at or above the median.

Figure 2 – Idaho Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(expressed in MAP Percentiles)



Note: Idaho's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of all 26 states reviewed in this study. Idaho's cut scores are consistently within 5 percentiles of the median.

Table 1 – Idaho Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	9	10	11	12	11	9
Mathematics	14	13	11	14	15	11

Note: This table ranks Idaho's cut scores relative to the cut scores of the other 25 states in the study. In third-grade reading, for example, Idaho ranks ninth out of 26, meaning that it surpassed 17 states and had lower cut scores than eight states.

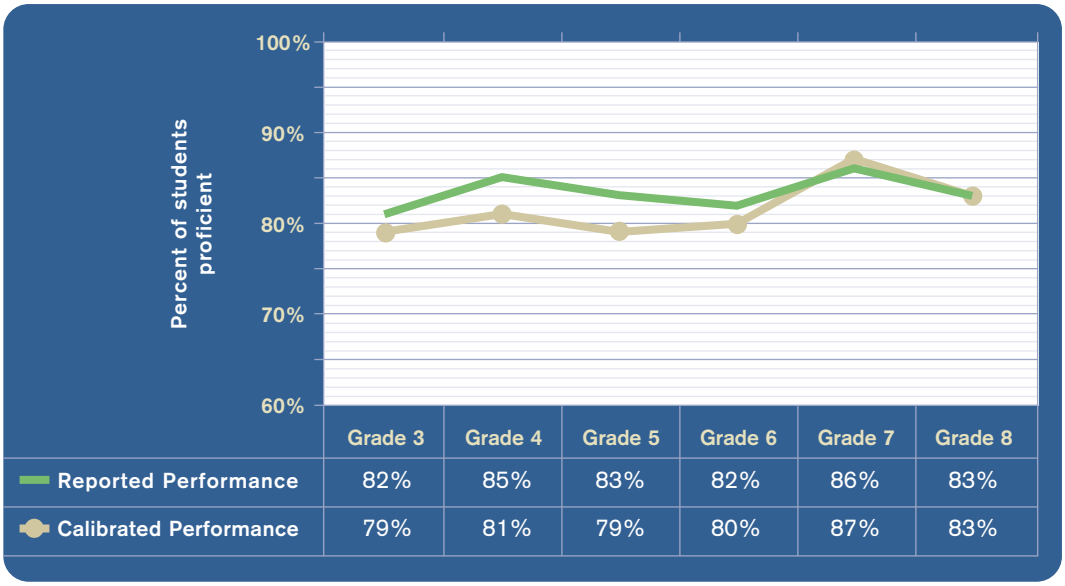
Part 2: Calibration across Grades*

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Idaho's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 indicated the relative difficulty of Idaho's reading and mathematics cut scores across grades, showing that, while the reading cut scores were fairly well calibrated, the math cut scores in the earlier grades were considerably easier than in the later grades. The following two figures show Idaho's reported performance in reading (Figure 3, page 76) and mathematics (Figure 4, page 77) on the state test and the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade 8 standard. Because the reading cut scores are fairly well calibrated across grades, Figure 3 shows little difference between the reported proficiency rates and the rates that would be expected if the cut scores were fully calibrated. Figure 4 shows that when differences in grade-to-grade difficulty of the mathematics cut score are removed, student performance is more consistent at all grades.

*Idaho is unique among the states in this report because it used NWEA's MAP as its official state assessment during the course of this study. This means that Idaho is the only state in which the cut scores were not derived by comparing the performance of a group of students on two instruments, but simply by reading Idaho's state test cut scores directly on the NWEA scale. It is impossible, therefore, to use the MAP as an independent ruler to determine whether Idaho's estimated cut scores inadvertently changed over time.

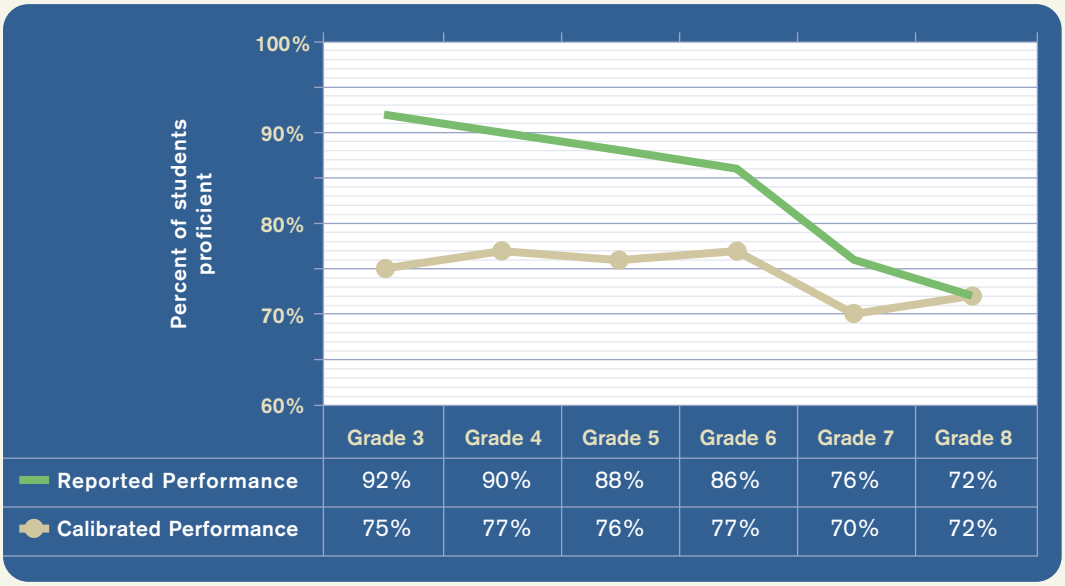
Figure 3 – Idaho Reading Performance as Reported and as Calibrated to the Grade 8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	82%	85%	83%	82%	86%	83%
Calibrated Performance	79%	81%	79%	80%	87%	83%

Note: This graphic shows, for example, that if Idaho's grade 3 reading cut score was as difficult as its grade 8 cut score, 79 percent of third graders would achieve the proficient level, rather than 82 percent, as was reported by the state.

Figure 4 – Idaho Mathematics Performance as Reported and as Calibrated to the Grade 8 Standard, 2006



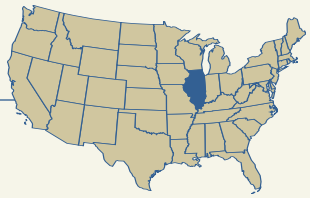
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	92%	90%	88%	86%	76%	72%
Calibrated Performance	75%	77%	76%	77%	70%	72%

Note: This graphic shows, for example, that if Idaho's grade 3 mathematics cut score was set at the same level of difficulty as its grade-8 cut score, 75% of third graders would achieve the proficient level, rather than 92%, as was reported by the state.

Policy Implications

When setting its cut scores for what students must know and be able to do in order to be considered proficient in reading and math, Idaho is about in the middle of the pack, at least compared with the other 25 states in this study. Unfortunately, these cut scores are not smoothly calibrated across grades, particularly in mathematics. Students who are proficient in third-grade mathematics are not necessarily on track to be proficient by the eighth grade. Idaho policymakers might consider raising their cut scores in the early grades so that parents and schools can be assured that young students scoring at the proficient level are truly prepared for success later in their education careers.

Illinois



Introduction

This study linked data from the spring 2003 and spring 2006 administrations of Illinois's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that most of Illinois's definitions of proficiency in reading and mathematics are lower than those of most of the other 25 states in this study. In other words, Illinois's tests are below average in terms of difficulty, especially in math.

Moreover, the level of difficulty generally declined from 2003 to 2006—the No Child Left Behind era—dramatically so in reading in grades 3 and 8, and in grade-8 math. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the Illinois test not being matched by learning gains on the Northwest Evaluation Association test. Nonetheless, Illinois's reading standards are still relatively higher for third grade than for eighth grade (taking into account the obvious differences in subject content and children's development). Consequently, the reading proficiency rates that the state reported for third grade actually underestimate the proportion of these students on track to meet the eighth-grade reading standards—even as Illinois's low cut scores in grade 8 might be masking performance problems at that level. Illinois's policymakers might take this opportunity to smooth and calibrate the state's reading standards, particularly in grade 8.

What We Studied: Illinois Standards Achievement Test (ISAT)

Illinois currently uses a spring assessment called the Illinois Standards Achievement Test (ISAT), which tests reading and math in grades 3 through 8, and science in grades 4 and 7. The current study analyzed reading and math results from a group of elementary and middle schools in which almost all students took both the state's assessment and MAP, using the spring 2003 and spring 2006 administrations of the two tests. (The methodology section of this report explains how performance on these two tests was compared.) These linked results were then used to estimate the scores on NWEA's scale that would be equivalent to the proficiency cut scores for each grade and subject on the Illinois State Assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.)

Part 1: How Difficult are Illinois's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

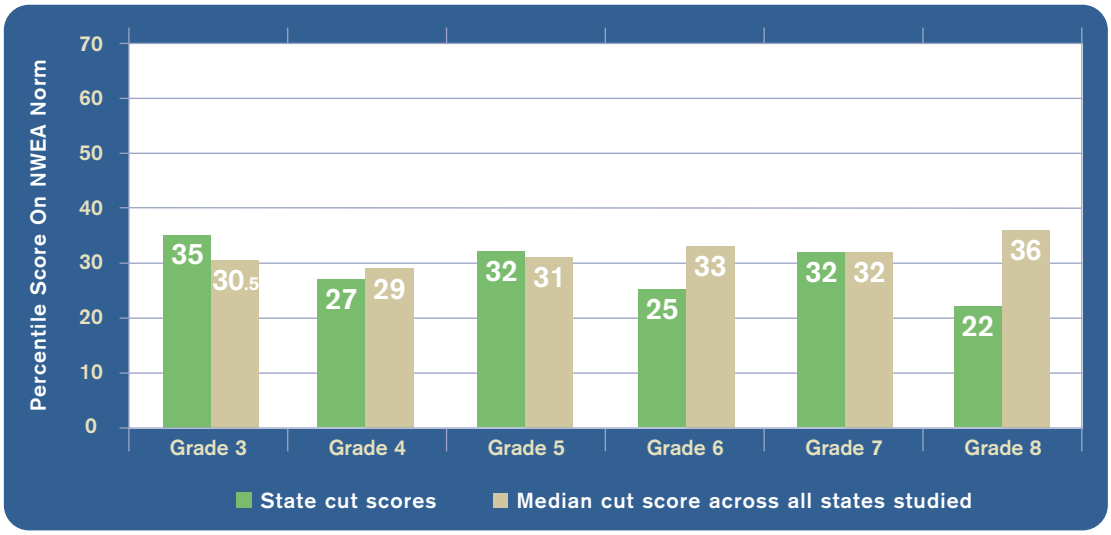
Applying that approach to this assignment, we evaluated the difficulty of Illinois's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Illinois standard on a test of equivalent difficulty. The two figures that follow show the difficulty of Illinois's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in spring 2006 in relation to the median cut scores for all the states in the study. The proficiency cut scores for **reading** in Illinois ranged between the 22nd and 35th percentiles of the NWEA norm group, with the third grade being most challenging—a rare circumstance among the states studied here. In **mathematics**, the proficiency cut scores fell to the 19th and 20th percentiles for the norm group except for fourth grade, where the cut score was less challenging.
Illinois's reading cut scores vary across grades, ranging from 14 points below the median to 4.5 points above the median, with eighth grade being conspicuously below the 26-state median.

In mathematics, cut scores for all grades are well below the median of the states studied.

Note, too, that Illinois's cut scores for reading are generally higher than for math. Thus, reported differences in achievement on the ISAT between reading and mathematics might be more a product of differences in cut scores than in actual student achievement. In other words, Illinois students might be performing better in reading and worse in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

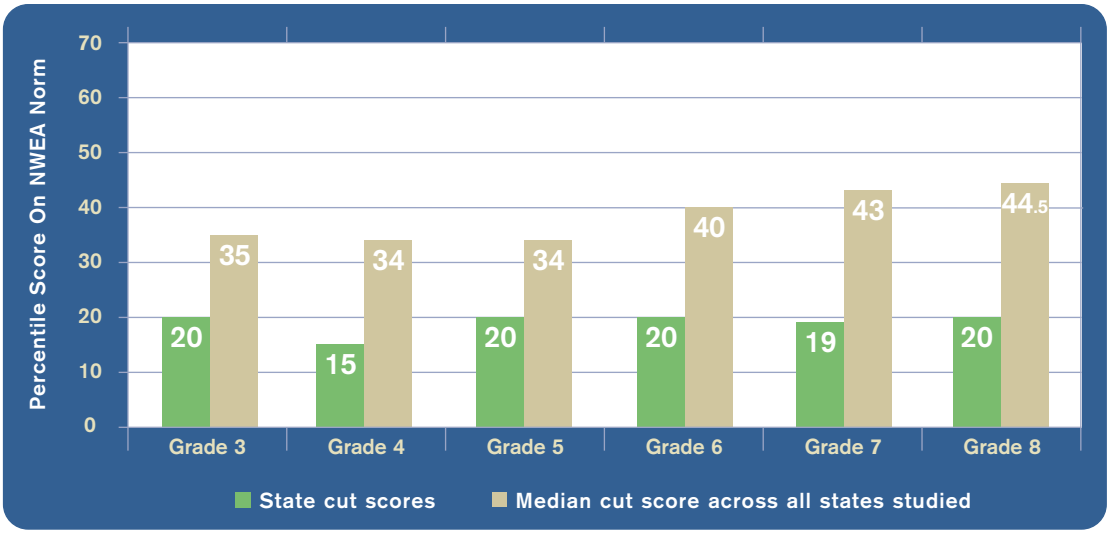
Another way of assessing difficulty is to ask how Illinois's proficiency cut scores rank relative to other states. Table 1 shows that Illinois's profieciency cut scores for reading rank in the mid- to upper third in difficulty (except in grades 6 and 8) among the 26 states studied for this report, while the cut scores for math rank in or near the lowest third in difficulty among the 26 states studied for this report.

Figure 1 – Illinois Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of other states reviewed in this study. Illinois ranks slightly above the median in both third and fifth grade, and its cut scores are at the median in seventh grade. Its eighth-grade cut score, however, is 14 percentile points below the median.

Figure 2 – Illinois Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles).



Note: Illinois's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. Illinois's cut scores in math are consistently 14 to 24.5 percentile points below the median.

Table 1 – Illinois Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	7	15	11	20	13	21
Mathematics	21	23	24	24	24	22

Note: This table ranks Illinois's cut scores relative to the cut scores of the other 25 states in the study, where 1 is highest and 26 is lowest.

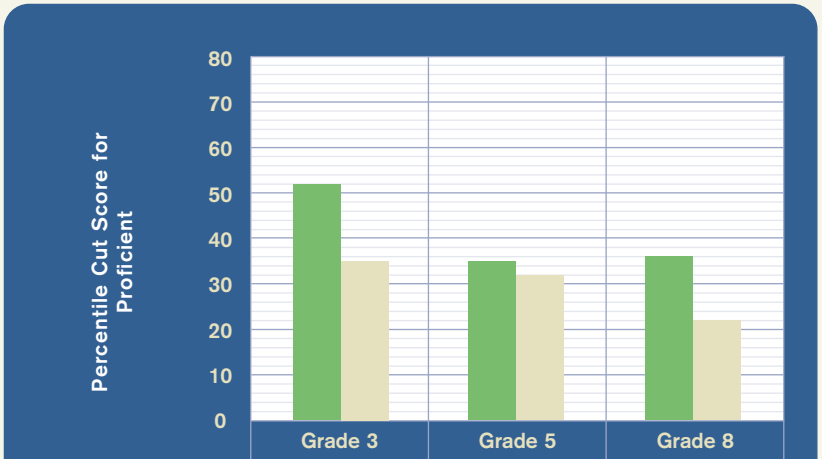
Part 2: Changes in Cut Scores over Time

In order to measure their consistency, Illinois's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2002-03 and 2005-06 school years. Cut score estimates for both years were available for grades 3, 5, and 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. This was the case for Illinois, which publicly changed its cut scores during the period studied.

Is it possible, then, to compare the proficiency scores between earlier administrations of Illinois's tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height to judge proficiency. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is slightly more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measure or scale used by the ISAT in 2003 and in 2006 can both be linked to the MAP test, which has remained consistent over time. Just as one can compare three feet with one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the ISAT in 2003 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty.

Figure 3 – Estimated Change in Illinois's Proficiency Cut Scores in Reading, 2003-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 5	Grade 8
Spring '03	52	35	36
Spring '06	35	32	22
Difference	-17	-3	-14

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third-grade students in 2003 had to score at the 52nd percentile of the NWEA norm group in order to be considered proficient, while in 2006 third graders had only to score at the 35th percentile to achieve proficiency. The change in grade 5 is within the margin of error (in other words, too small to be considered substantive).

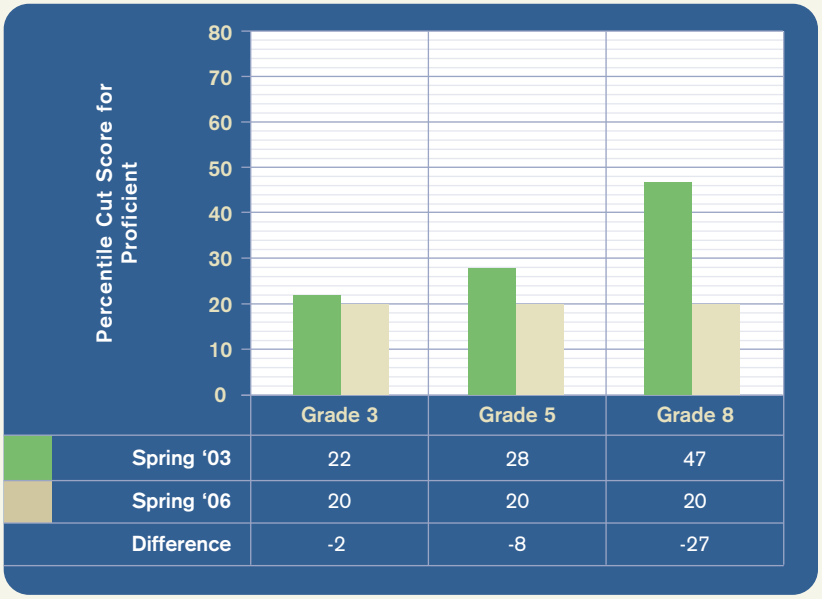
For **reading**, we found a decrease in Illinois's estimated proficiency cut scores in grades three and eight over this three-year period (Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, these changes would likely yield increases in the third-grade reading proficiency rate by 17 percent and in the eighth-grade reading proficiency rate by 14 percent. (Illinois reported a 9 point gain for third graders and a 16 point gain for eighth graders over this period.)

Analyses of Illinois's estimated **mathematics** proficiency cut scores indicate a decrease in grades 5 and 8 over this three-year period (Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, this would likely yield increased proficiency rates of 8 percent and 27 percent, respectively. (Illinois reported a 10-point gain for fifth graders and a 25-point gain for eighth graders over this period.)

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figure 4 – Estimated Differences in Illinois's Proficiency Cut Scores in Mathematics, 2003-2006 (Expressed in MAP Percentiles).

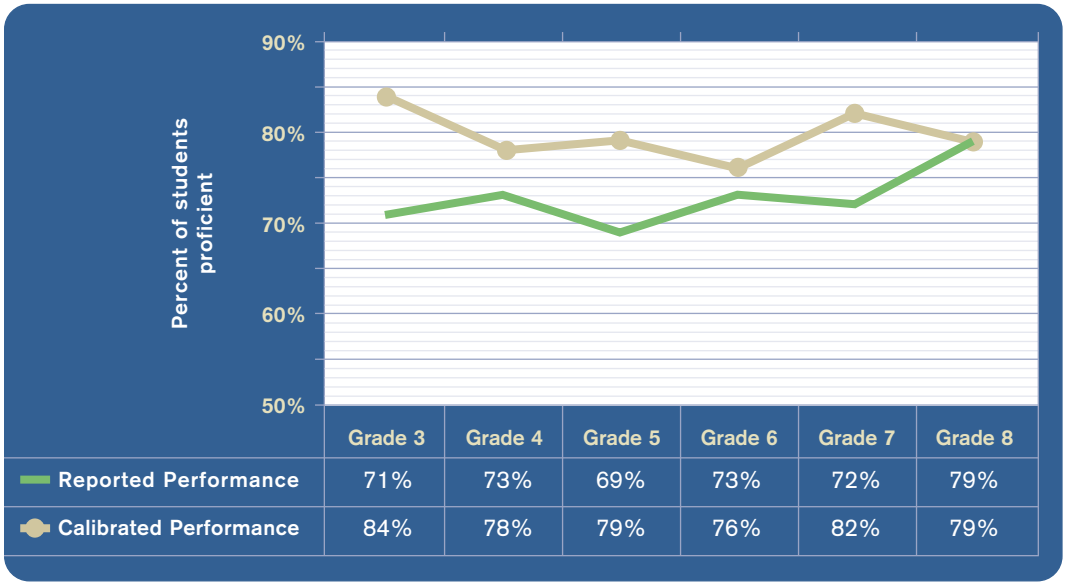


	Grade 3	Grade 5	Grade 8
Spring '03	22	28	47
Spring '06	20	20	20
Difference	-2	-8	-27

Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, eighth-grade students in 2003 had to score at the 47th percentile of the NWEA norm group in order to be considered proficient, while in 2006 eighth graders only had to score at the 20th percentile of the NWEA norm group to achieve proficiency. The change in grades 3 was within the margin of error (in other words, too small to be considered substantive).

Examining Illinois's cut scores, we find that they are not well calibrated across grades. Figure 1 showed that Illinois's reading proficiency cut scores in third grade are relatively more challenging than in eighth grade. Figure 2 showed that the math proficiency cut score is fairly consistent across the grades. The two figures that follow show Illinois's reported performance on its state test in reading (Figure 5) and mathematics (Figure 6) compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut scores are removed, it becomes clear that the percentage of elementary and middle school students who are on track to meet the eighth-grade reading proficiency cut scores is actually higher than what was reported by the state.

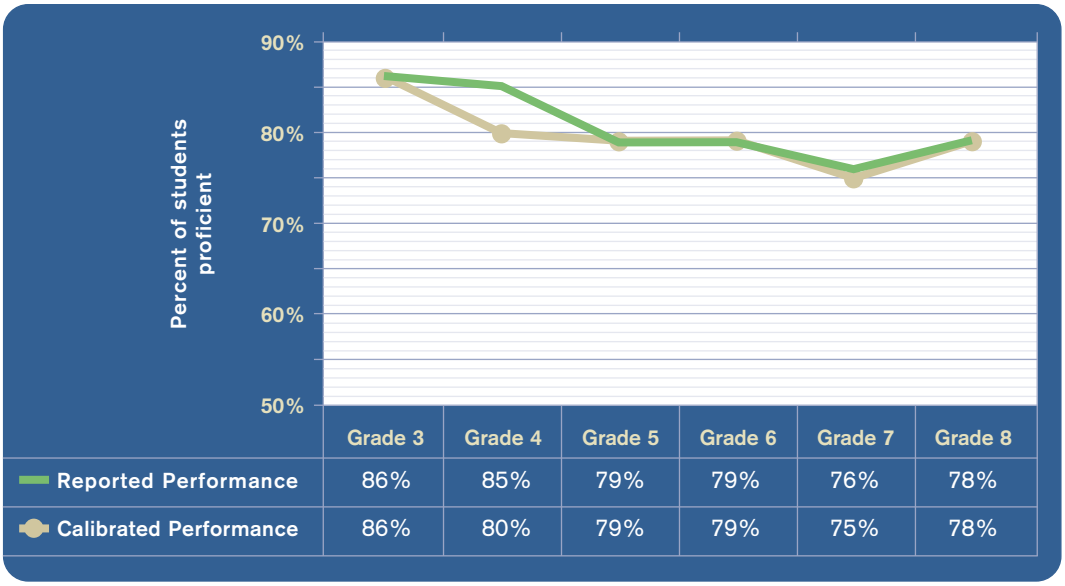
Figure 5 – Illinois Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	71%	73%	69%	73%	72%	79%
Calibrated Performance	84%	78%	79%	76%	82%	79%

Note: This graphic shows, for example, that if Illinois's grade-3 reading standard were set at the same level of difficulty as its grade-8 standard, 84 percent of third graders would achieve the proficient level, rather than 71 percent, as reported by the state.

Figure 6 – Illinois Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



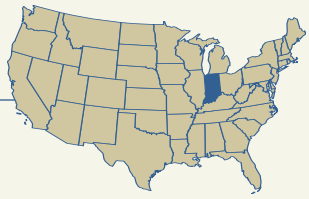
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	86%	85%	79%	79%	76%	78%
Calibrated Performance	86%	80%	79%	79%	75%	78%

Note: This graphic shows, for example, that if Illinois's grade-4 mathematics standard were set at the same level of difficulty as its grade-8 standard, 80 percent of fourth graders would achieve the proficient level, rather than 85 percent, as was reported by the state. Fourth grade aside, it appears that Illinois math standards are fairly well calibrated from grade to grade.

Policy Implications

Illinois's proficiency cut scores are relatively low for math and about average for reading, compared with the other 25 states in the study. This finding is fairly consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, particularly for reading in the higher grades (although not as much for math). Reading and math standards have generally decreased between 2003 and 2006, dramatically in some grades. Moreover, Illinois's expectations for reading proficiency are not smoothly calibrated across grades; Illinois's third-grade proficiency rates actually underestimate the proportion of students who are on track to meet the eighth-grade requirements. Illinois policymakers might consider raising all of their cut scores, but especially those at the eighth-grade level.

Indiana



Introduction

This study linked data from the 2002 and 2006 administrations of Indiana's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Indiana's definitions of "proficiency" in reading and mathematics are somewhat below the standards set by the other 25 states in this study. In other words, Indiana's tests are a bit below average in terms of difficulty.

The difficulty of Indiana proficiency cut scores decreased somewhat from 2002 to 2006—the No Child Left Behind era—although not for all grades. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the Indiana test not being matched by learning gains on the Northwest Evaluation Association test. One striking finding is that Indiana's reading cut scores are easier for third-grade students than for eighth-grade pupils (taking into account the obvious differences in subject content and children's development). State policymakers might consider adjusting their reading cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Indiana Statewide Testing for Educational Progress-Plus (ISTEP+)

Indiana currently uses an assessment called the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+), which tests English/language arts and math in grades 3-10, and science in grades 5 and 7. This test has been in use since fall 2003, replacing the Indiana Statewide Testing for Educational Progress (ISTEP). The current study linked results from fall 2002 ISTEP administrations and fall 2006 ISTEP+ administrations to a common scale also administered in the 2002 and 2006 school years.

To determine the difficulty of Indiana's proficiency cut scores, we linked reading and math data from Indiana's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of schools in which almost all students took both the state assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Indiana's Definitions of Proficiency in Reading and Math?

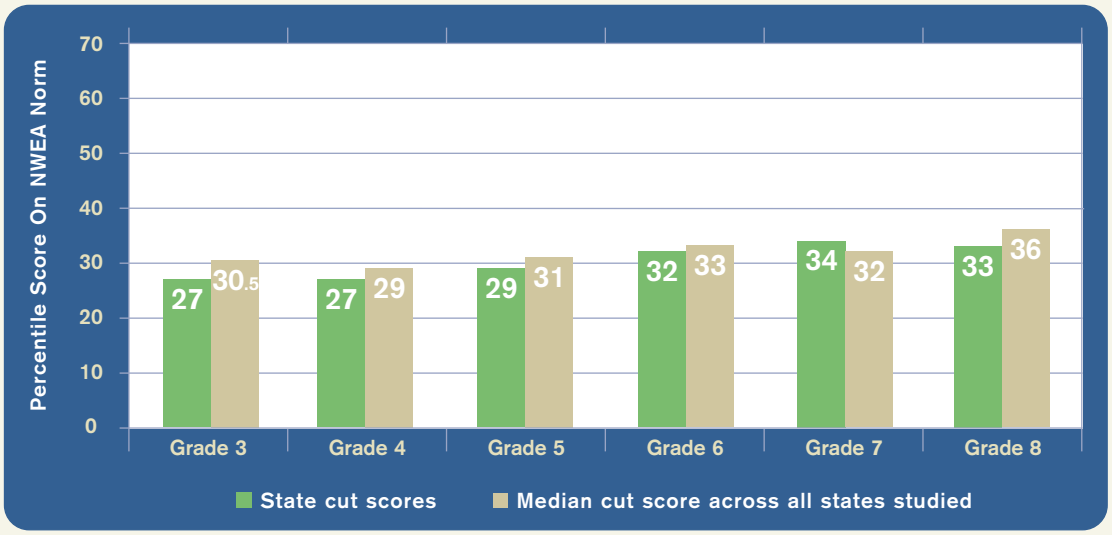
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high-jump bar is easy to jump over? We know because if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high-jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this inquiry, we evaluated the difficulty of Indiana's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Indiana cut score on a test of equivalent difficulty. The following two figures show the difficulty of Indiana's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Indiana ranged between the 27th and 34th percentiles for the norm group, with the seventh grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 26th and 35th percentiles for the norm group, with third grade being most challenging.

For most grade levels, Indiana's cut scores in reading and mathematics are consistently near the median level among the states studied. Math cut scores for grades six through eight, however, are well below the median levels of difficulty.

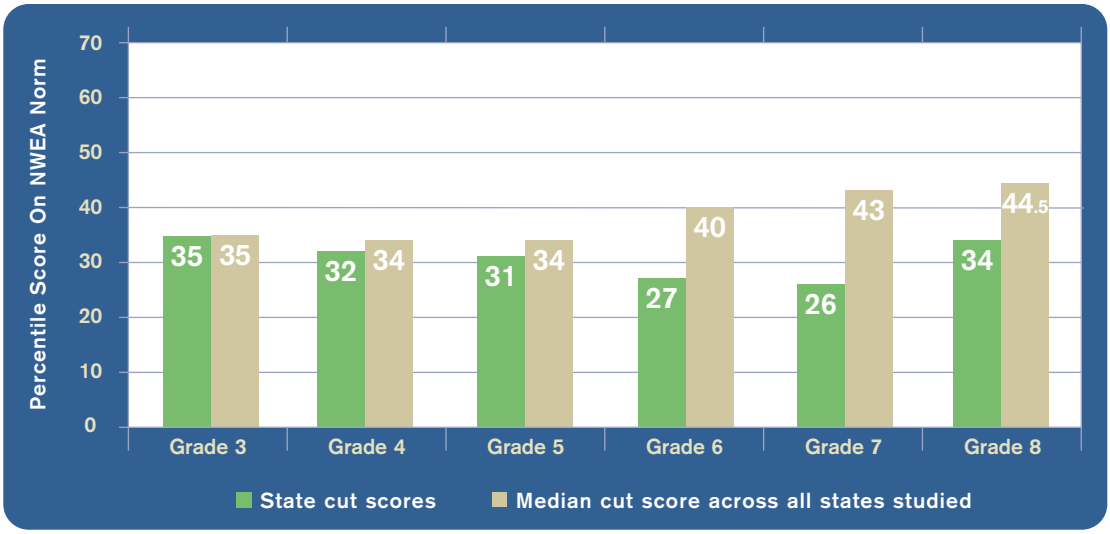
Another way of assessing difficulty is to evaluate how Indiana's proficiency cut scores rank relative to other states. Table 1 shows that Indiana cut scores generally rank in the mid- or bottom third in difficulty among the 26 states studied for this report.

Figure 1 – Indiana Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of other states reviewed in this study. Only in seventh grade does Indiana's cut score reach above the median. Grades 3-6 and grade 8 scores are 1 to 3.5 percentile points below the median.

Figure 2 – Indiana Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: Indiana's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. Only in third grade does Indiana's math cut score reach the median; otherwise, it is 2 to 17 percentile points below.

Table 1 – Indiana's Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	15	15	16	14	12	14
Mathematics	13	16	17	21	22	17

Note: This table ranks Indiana's cut scores relative to the cut scores of the other 25 states in the study, where 1 is highest and 26 is lowest.

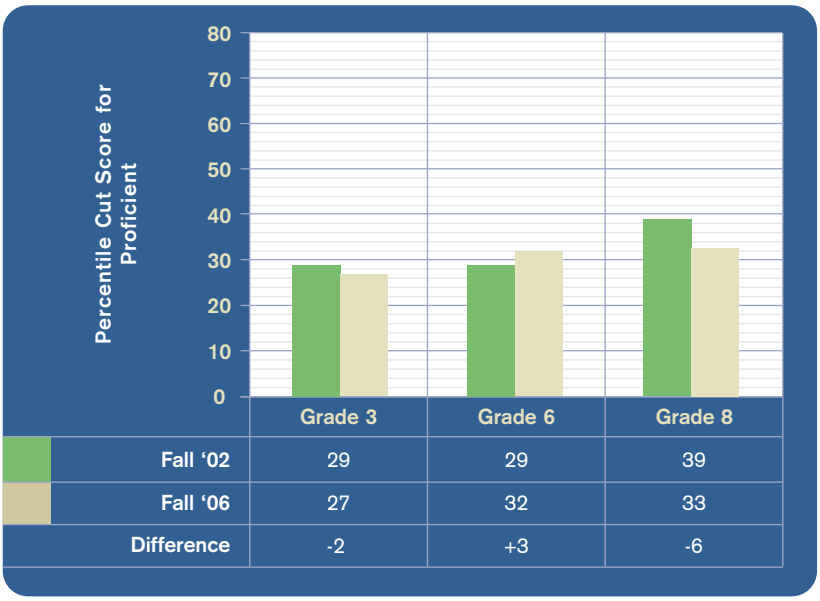
Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Indiana's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2002 and 2006 school years. Cut score estimates for both years were available in both reading and mathematics for grades 3, 6, and 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the assessments used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed.

Is it possible, then, to compare the proficiency scores between earlier administrations of Indiana's tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The ISTEP in

Figure 3 – Estimated Differences in Indiana's Proficiency Cut Scores in Reading, 2002-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 6	Grade 8
Fall '02	29	29	39
Fall '06	27	32	33
Difference	-2	+3	-6

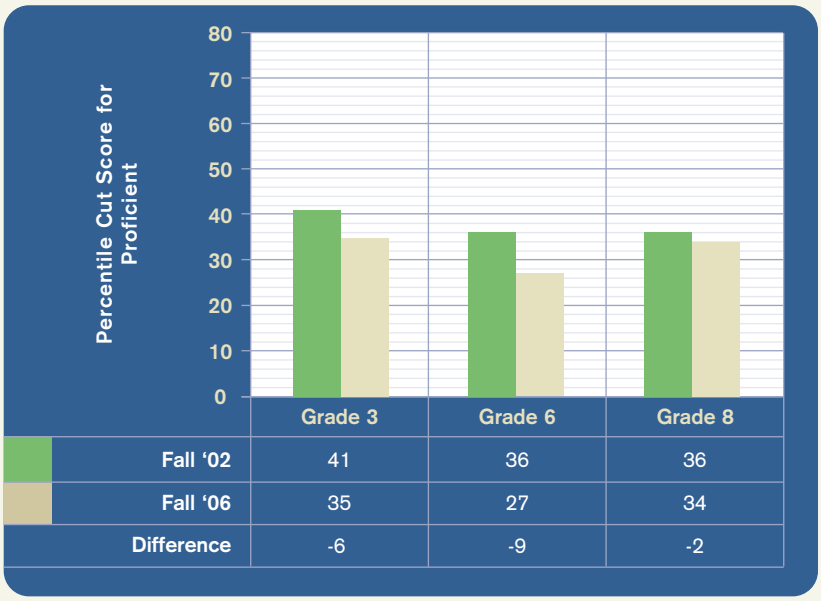
Note: This graphic shows whether the difficulty of achieving proficiency in reading has changed. For example, eighth-grade students in 2002 had to score at the 39th percentile of the NWEA norm group in order to be considered proficient, while in 2006 eighth graders had only to score at the 33rd percentile of the NWEA norm group to achieve proficiency, although this change is not substantive. The changes in grades 3, 6, and 8 were within the margin of error (in other words, too small to be considered substantive).

2002 and the ISTEP+ in 2006 can both be linked to the MAP, which has remained consistent over time. This allows us to estimate whether the ISTEP+ in 2006 was easier to pass, harder, or about the same as the ISTEP in 2002. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass Indiana's assessments in 2002 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty.

In **reading**, no substantive differences are visible in grades 3, 6, and 8 (the observed changes were smaller than the margin of error for the estimate, see Figure 3).

Indiana's estimated **mathematics** cut scores decreased moderately for sixth grade (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect to see a 9 percent increase for sixth graders. (Indiana reported a 12-point gain for sixth graders over this period.)

Figure 4 – Estimated Difference in Indiana's Proficiency Cut Scores in Mathematics, 2002-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 6	Grade 8
Fall '02	41	36	36
Fall '06	35	27	34
Difference	-6	-9	-2

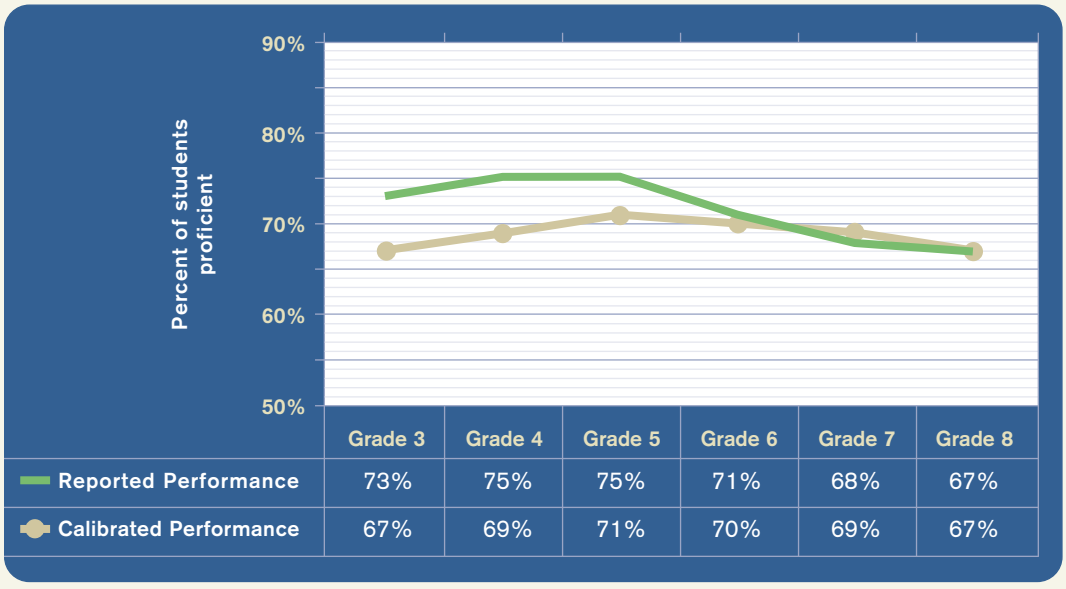
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, sixth-grade students in 2002 had to score at the 36th percentile of the NWEA norm group in order to be considered proficient, while in 2006 third graders only had to score at the 27th percentile of the NWEA norm group to achieve proficiency. The changes in grades 3 and 8 were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Indiana's cut scores, we find that they are not well calibrated across grades. Figure 1 showed that Indiana's upper grade cut scores in reading in 2006 were more challenging than the cut scores in the lower grades. A different pattern emerged in mathematics, with the cut scores at third and eighth grades being more challenging than the grades in between (see Figure 2). The two figures that follow show Indiana's reported performance on its state test in reading (Figure 5) and mathematics (Figure 6), compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the eighth-grade standard. When differences in grade-to-grade difficulty of the cut scores are removed, student performance in both reading and math is more consistent at all grades. This would lead to the conclusion that the higher rates of proficiency that the state has reported for elementary school students in reading are somewhat misleading.

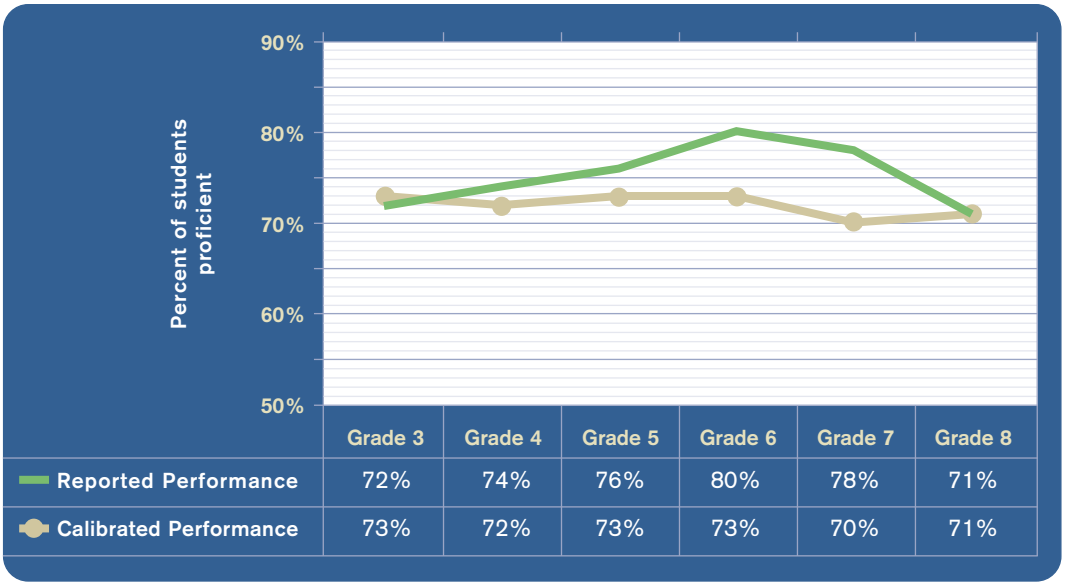
Figure 5 – Indiana Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	73%	75%	75%	71%	68%	67%
Calibrated Performance	67%	69%	71%	70%	69%	67%

Note: This graphic shows, for example, that if Indiana's grade-3 reading standard were set at the same level of difficulty as its grade-8 standard, 67 percent of third graders would achieve the proficient level, rather than 73 percent, as reported by the state.

Figure 6 – Indiana Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



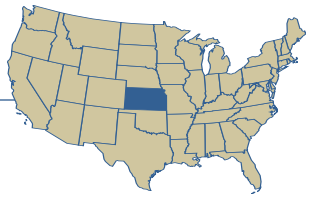
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	72%	74%	76%	80%	78%	71%
Calibrated Performance	73%	72%	73%	73%	70%	71%

Note: This graphic shows, for example, that if Indiana's grade-7 mathematics cut score were set at the same level of difficulty as its grade-8 standard, 70 percent of seventh graders would achieve the proficient level, rather than the 78 percent reported by the state.

Policy Implications

When setting its cut scores for what it takes for a student to be considered proficient in reading and math, Indiana is slightly below average, at least compared with the other 25 states in this study. (This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Indiana's standards to be about average in the distribution of all states studied.) Indiana's cut scores have remained fairly constant over the past several years, although eighth-grade reading and third- and sixth-grade math standards have eased. However, Indiana's expectations are imperfectly calibrated across grades; students who are proficient in third-grade reading, in particular, are not necessarily on track to be proficient by the eighth grade. Indiana policymakers might consider adjusting their reading cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Kansas



Introduction

This study linked data from the 2006 administration of Kansas's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Kansas's definitions of "proficiency" in reading and mathematics are relatively consistent with the standards set by the other 25 states in this study. In other words, Kansas's tests are about average in terms of difficulty.

Like many states, however, Kansas's math proficiency cut scores are easier in the earlier grades than in the later grades (taking into account the obvious differences in subject content and children's development). Therefore, the reported proficiency rates may overestimate the proportion of third-grade students who are actually on track to be proficient in eighth-grade mathematics. Moreover, Kansas's reading cut scores are generally easier than the state's corresponding math cut scores for a given grade. State policymakers might consider adjusting their math cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

What We Studied: Kansas Assessment System

The current Kansas Assessment tests mathematics in students in grades 3-8, and grade 10, and reading in students in grades 3-8, and grade 11. This study linked data from spring 2006 to a common scale also administered in the 2006 school year.

To determine the difficulty of Kansas's proficiency cut scores, we linked data from state tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of schools in which almost all students took both the Kansas Assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Kansas's Definitions of Proficiency in Reading and Math?

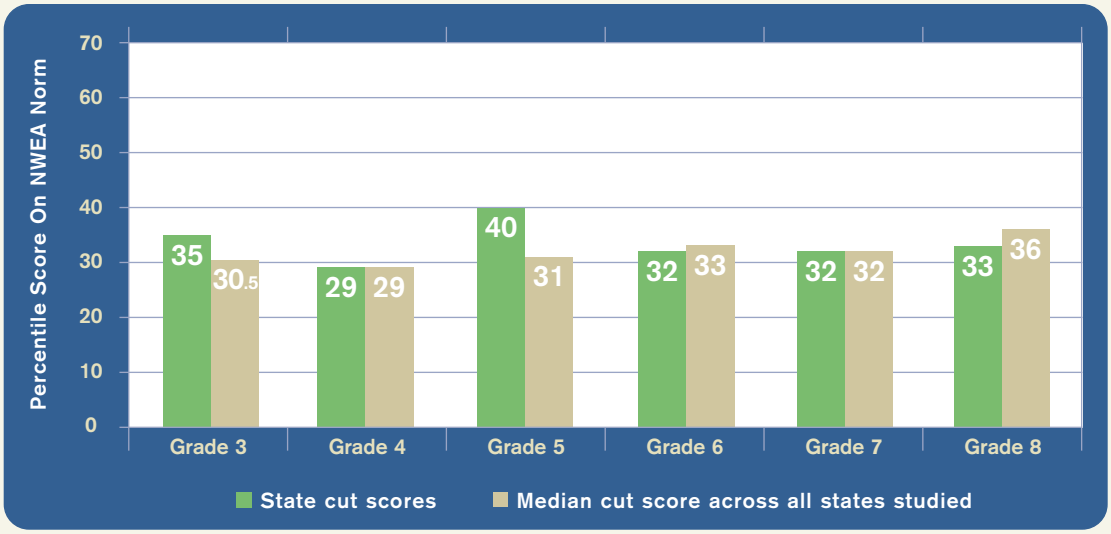
One way to assess the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to leap? We know because if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? We know because only one (or perhaps none) of those same 100 individuals would successfully meet that level of challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that concept to this analysis, we evaluated the difficulty of the Kansas proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the cut score on a test of equivalent difficulty. The following two figures show the difficulty of Kansas proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Kansas ranged between the 29th and 40th percentiles of the norm group, with the fifth grade being most challenging. In **mathematics**, the cut scores ranged between the 30th and 45th percentiles with the seventh grade being most challenging.

With a few exceptions, Kansas's cut scores in reading and math are near the median level of difficulty of all 26 states in this study. Note, though, that Kansas's reading cut scores are generally easier than the corresponding math cut score for a given grade. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Kansas students might be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

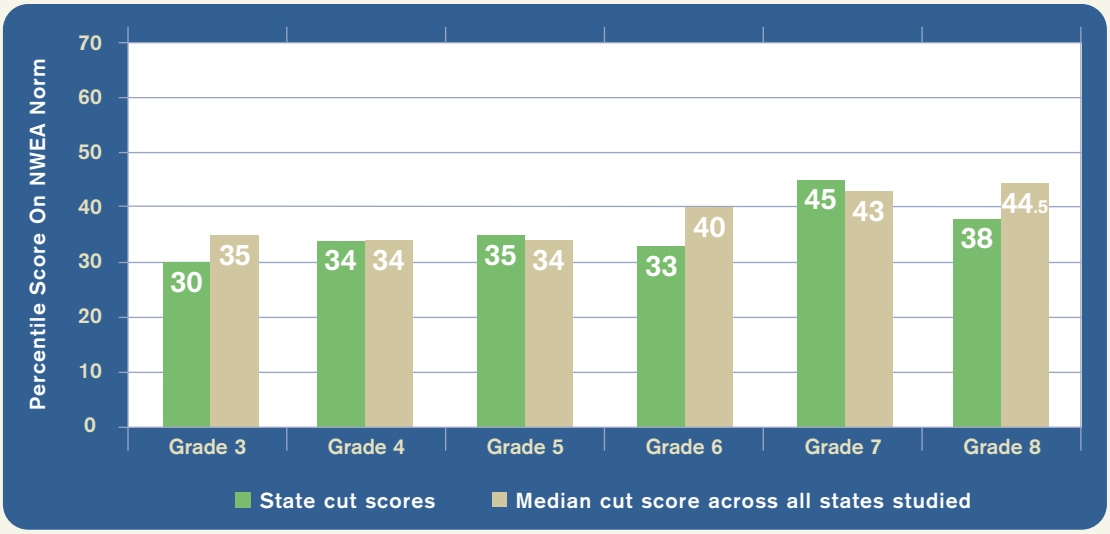
Another way of assessing difficulty is to evaluate how Kansas's proficiency cut scores rank relative to other states. Table 1 shows that the Kansas cut scores generally rank in the middle third in difficulty among the 26 states studied for this report.

Figure 1 – Kansas Reading Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in 2005 MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of other states reviewed in this study. Kansas's cut scores are generally near the median except in grades 3 and 5, which are respectively 4.5 and 9 percentile points above the median.

Figure 2 – Kansas Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in 2005 MAP Percentiles)



Note: Kansas's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of all 26 states reviewed in this study. The cut scores are close to the median in grades 4, 5, and 7, but slip below in grades 3, 6, and 8.

Table 1 – Kansas Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	7	13	6	14	13	14
Mathematics	14	13	11	18	8	14

Note: This table ranks Kansas's cut scores relative to the cut scores of the other 25 states in the study, where 1 is highest and 26 is lowest.

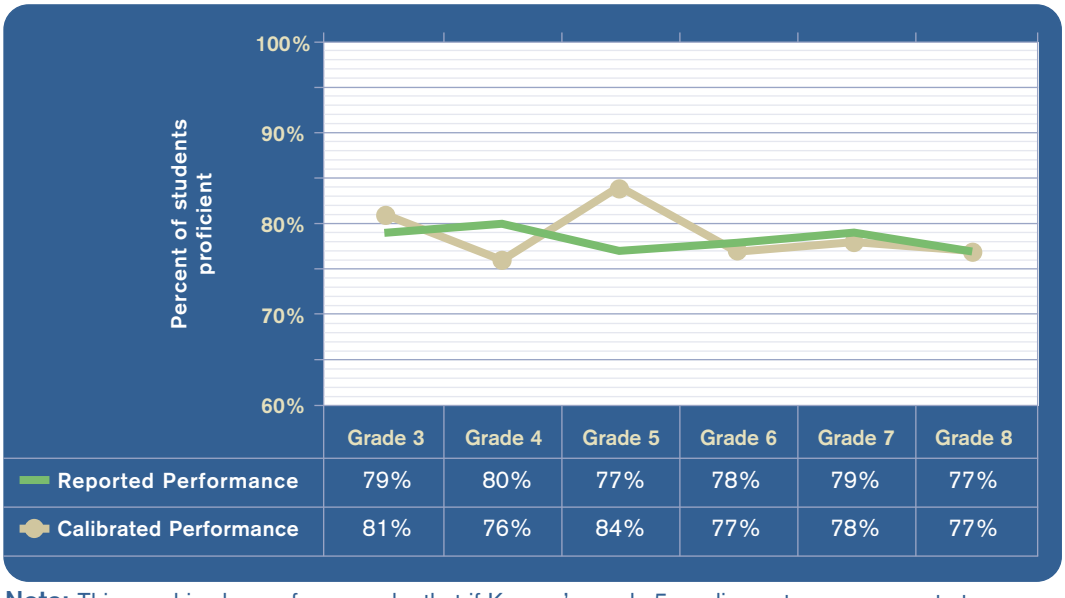
Part 2: Calibration across Grades*

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Kansas's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 above illustrated the relative difficulties of the Kansas's reading and math cut scores, showing how the mathematics proficiency cut scores for the lower grades were somewhat less difficult than for the higher grades. The two figures that follow show Kansas's reported performance in reading (Figure 3) and mathematics (Figure 4) on the state test, compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade 8 standard. This has little effect in reading but when the differences in grade-to-grade difficulty of the cut score are removed in math, student performance changes, suggesting that the higher rates of mathematics proficiency that the state has reported for elementary school students are somewhat misleading.

*Kansas was one of seven states in this study for which cut score estimates could be determined for only one time period. Therefore, it was not possible to examine whether the state's cut scores have changed over time.

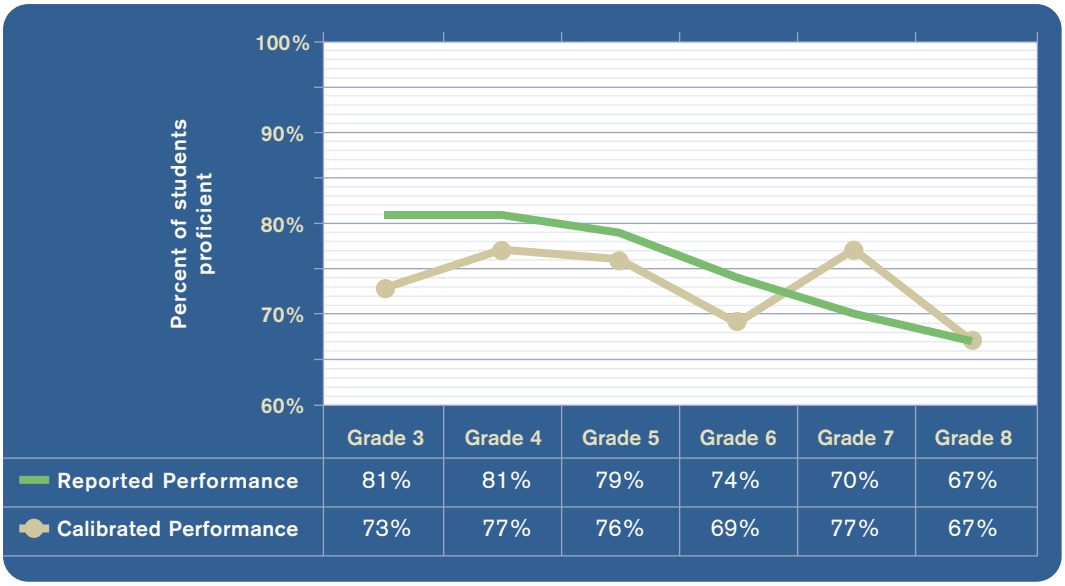
Figure 3 – Kansas Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	79%	80%	77%	78%	79%	77%
Calibrated Performance	81%	76%	84%	77%	78%	77%

Note: This graphic shows, for example, that if Kansas's grade-5 reading cut score was set at the same level of difficulty as its grade-8 cut score, 84 percent of fifth graders would achieve the proficient level, rather than 77 percent, as was reported by the state.

Figure 4 – Kansas Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



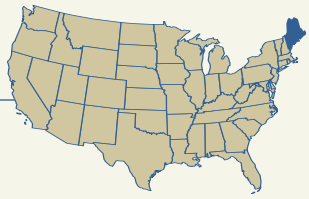
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	81%	81%	79%	74%	70%	67%
Calibrated Performance	73%	77%	76%	69%	77%	67%

Note: This graphic shows, for example, that if Kansas's grade-3 mathematics cut score was set at the same level of difficulty as its grade-8 standard, 73 percent of third graders would achieve the proficient level, rather than 81 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what it takes for a student to be considered proficient in reading and math, Kansas is generally near the middle of the pack, compared to the other 25 states in this study. This finding is fairly consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales,* which found Kansas's standards to be in the middle-third of the distribution of all states studied in grade-8 reading. Kansas's math proficiency cut scores are not smoothly calibrated across grades, however; students who are proficient in third-grade math are not necessarily on track to be proficient by the eighth grade. Kansas policymakers might consider adjusting their math cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

Maine



Introduction

This study linked data from the 2004 and 2006 administrations of Maine's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Maine's definitions of "proficiency" in reading and mathematics are relatively difficult compared with the standards set by the other 25 states in this study. In other words, Maine's tests are above average in terms of difficulty.

Yet the difficulty level of Maine's tests decreased dramatically from 2004 to 2006—the No Child Left Behind era. This is not a surprise, as Maine adopted a new scale for both the reading and math tests for the 2005-06 academic school year, and publicly reported lowering the cut scores on those tests.

Not well known, however, is that Maine's cut scores in reading and math are easier for third-grade students than for eighth-grade pupils (taking into account the differences in subject content and children's development). Plus, as is true for the majority of states studied, Maine's cut scores for reading are lower than those for mathematics. Maine policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

What We Studied: Maine Educational Assessment (MEA)

Maine currently uses an assessment called the Maine Educational Assessment (MEA) which tests reading and mathematics in grades 3 to 8, writing in grades 5 and 8, and science in grades 4 and 8. The current study linked reading and math results from spring 2004 and spring 2006 MEA administrations to a common scale also administered in the 2004 and 2006 school years. Sample sizes for the 2004 testing season were not sufficiently large to meet the inclusion criteria for the national findings sections of the overall report (at least 700 students per grade, whereas in the Maine 2004 sample, only about 400 per grade were available for math, and about 300 for reading). Consequently, the findings in section 2 of this Maine report are not included in the national report. They are included in the state report for informational purposes, but because of the small sample sizes upon which they are based, they should be interpreted with caution.

To determine the difficulty of Maine's proficiency cut scores, we linked data from Maine's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Maine's Definitions of Proficiency in Reading and Math?

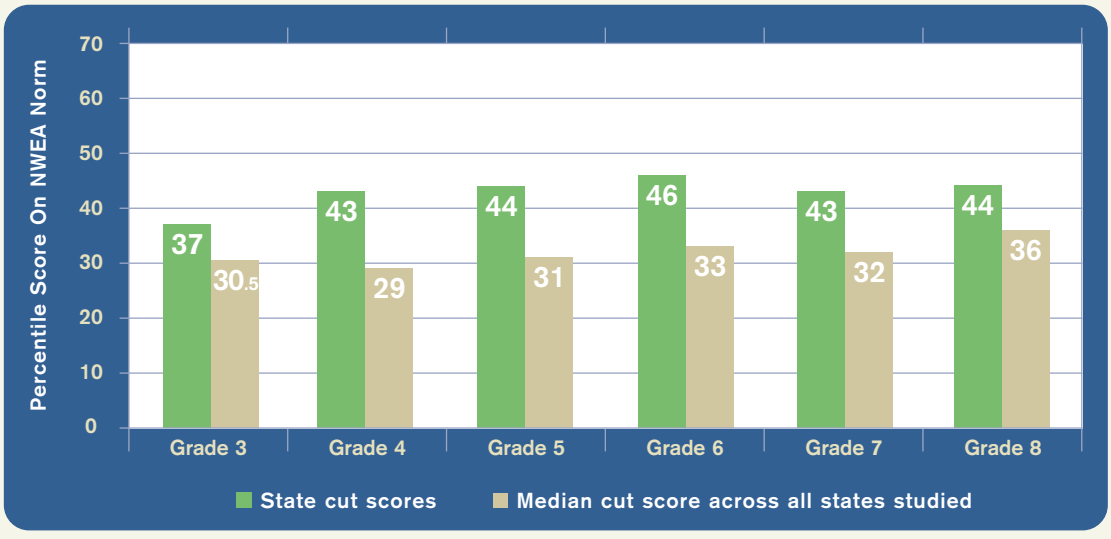
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of Maine's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Maine cut score on a test of equivalent difficulty. The following two figures show the difficulty of Maine's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Maine ranged between the 37th and 46th percentiles in the norm group, with the sixth-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 43rd and 54th percentiles with seventh grade being most challenging.

Maine's cut scores in both reading and mathematics are consistently above the median difficulty level among the states studied. In other words, Maine's tests are harder to pass than the average state test. Note, though, that Maine's cut scores for reading are lower than for math. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. Maine students might be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

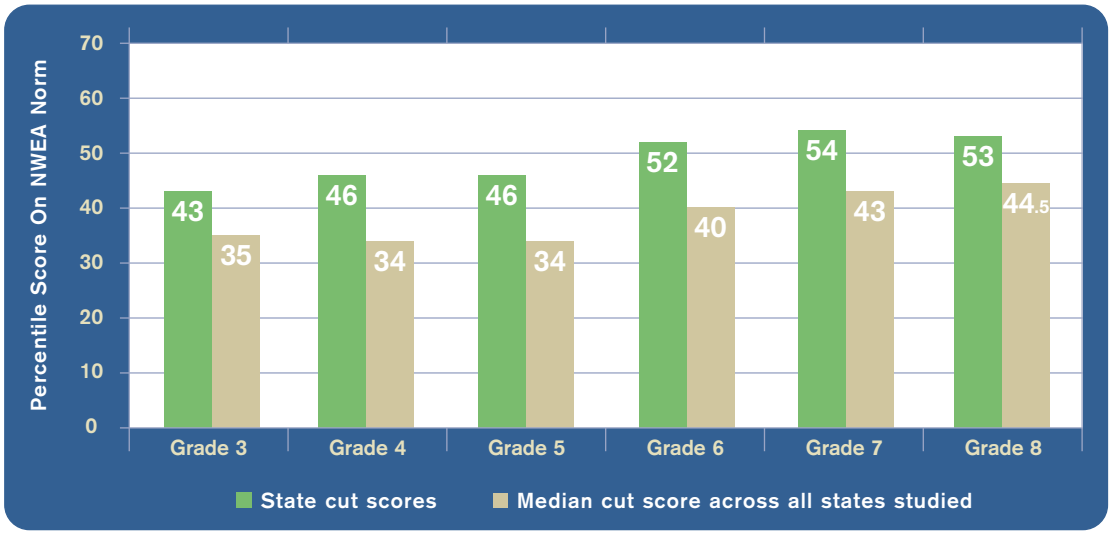
Another way of assessing difficulty is to evaluate how Maine's proficiency cut scores rank relative to other states. Table 1 shows that the Maine cut scores generally rank in the upper third in difficulty among the 26 states studied for this report. Its reading cut scores are particularly high, ranking third among the states in grades 4 and 6.

Figure 1 – Maine Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of other states reviewed in this study. Maine's cut scores are consistently above the median.

Figure 2 – Maine Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: Maine's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. Maine's cut scores are consistently above the median.

Table 1 – Maine Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	5	3	5	3	5	6
Mathematics	6	5	8	6	6	6

Note: This table ranks Maine's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

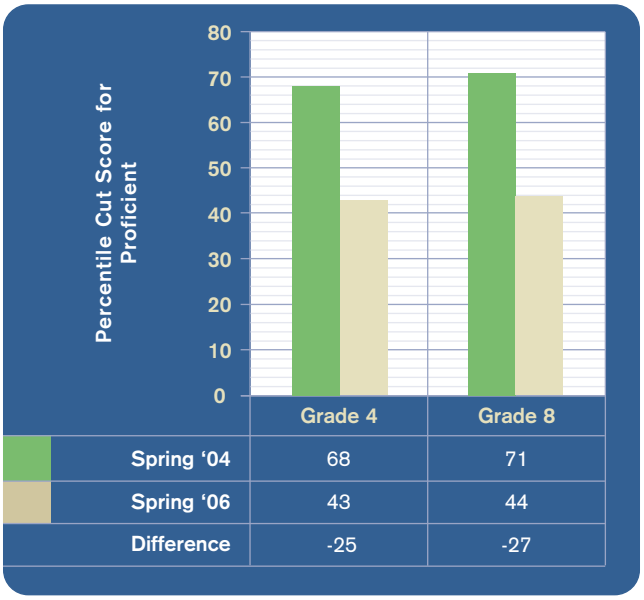
Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Maine's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2004 and 2006 school years. Cut score estimates for reading and mathematics were available for both years for grades 4 and 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. This occurred in Maine in the 2005-06 academic year, when the state adopted new scales and publicly lowered cut scores for both the reading and math tests.

Is it possible, then, to compare the proficiency scores between earlier administrations of Maine's tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. MEA in 2004 and MEA in 2006 can both be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to a meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the MEA in 2004 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty—and whether those changes are consistent with what the state reported to the public.

Figure 3 – Estimated Differences in Maine's Proficiency Cut Scores in Reading, 2004-2006 (Expressed in MAP Percentiles)



	Grade 4	Grade 8
Spring '04	68	71
Spring '06	43	44
Difference	-25	-27

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, fourth-grade students in 2004 had to score at the 68th percentile with respect to the NWEA norm group in order to be considered proficient, while by 2006 fourth graders had only to score at the 43rd percentile to achieve proficiency.

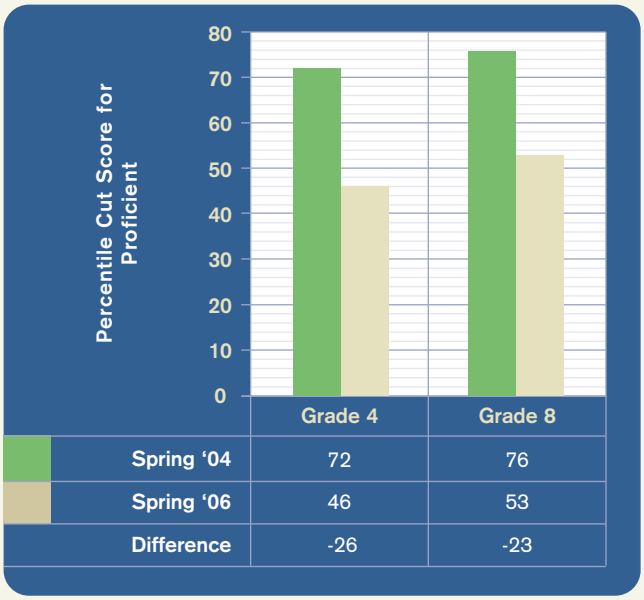
The sample size for the Maine 2004 testing season was not sufficiently large to meet the inclusion criteria for this study (i.e., estimates were based on fewer than 700 students per grade). Consequently, the discussions of "differences over time" that appear in the national sections of the overall report do not include Maine. These findings are reported for informational purposes, and should be interpreted with caution.

Despite the fact (see Figures 1 and 2) that Maine's 2006 cut scores were among the more challenging in the country, the state's estimated **reading** cut scores declined over this period in fourth and eighth grade (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the fourth-grade reading proficiency rate in 2006 to be 25 percent higher than in 2004. Similarly, one would expect eighth-grade reading proficiency rates to increase by 27 percent. (Maine reported a 11 point gain for fourth graders and a 22 point gain for eighth graders over this period.)

In **mathematics**, Maine's estimated cut scores show the same pattern as in reading, with visible erosion in the difficulty of the fourth- and eighth-grade cut scores (see Figure 4. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, these decreases would likely yield 26 percent and 23 percent increases in the reported math proficiency rates for fourth and eighth-grade students, respectively. (Maine reported a 27 point gain for fourth graders and a 23 point gain for eighth graders over this period.)

Thus, one could fairly say that Maine's reading and math tests were much easier to pass in 2006 than in 2004. It is important to note, however, that even with these decreases in difficulty, Maine's tests are still harder to "pass" than those of many other states in the study.

Figure 4 – Estimated Differences in Maine's Proficiency Cut Scores in Mathematics, 2004-2006 (Expressed in MAP Percentiles)



	Grade 4	Grade 8
Spring '04	72	76
Spring '06	46	53
Difference	-26	-23

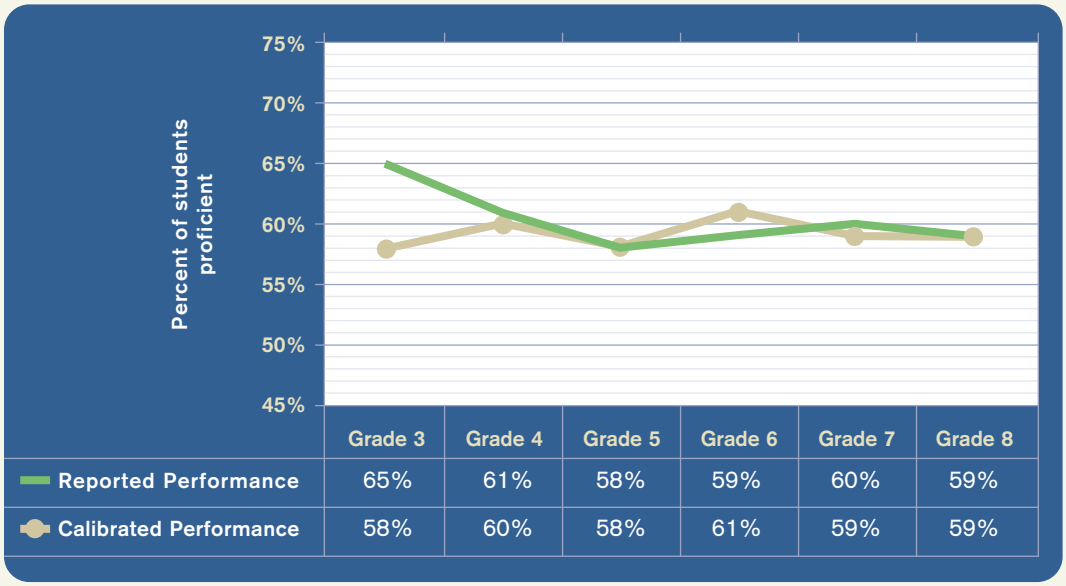
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, fourth-grade students in 2004 had to score at the 72nd percentile nationally in order to be considered proficient, while by 2006 fourth graders only had to score at the 46th percentile to achieve proficiency.

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Maine's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 above showed that Maine's upper-grade cut scores in reading and mathematics in 2006 were somewhat more challenging than the cut scores in the lower grades, particularly grade 3. The two figures that follow show Maine's reported performance on its state tests in reading (Figure 5) and mathematics (Figure 6), compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades, especially in math. This would lead to the conclusion that the higher rates of mathematics proficiency that the state has reported for elementary school students are somewhat misleading.

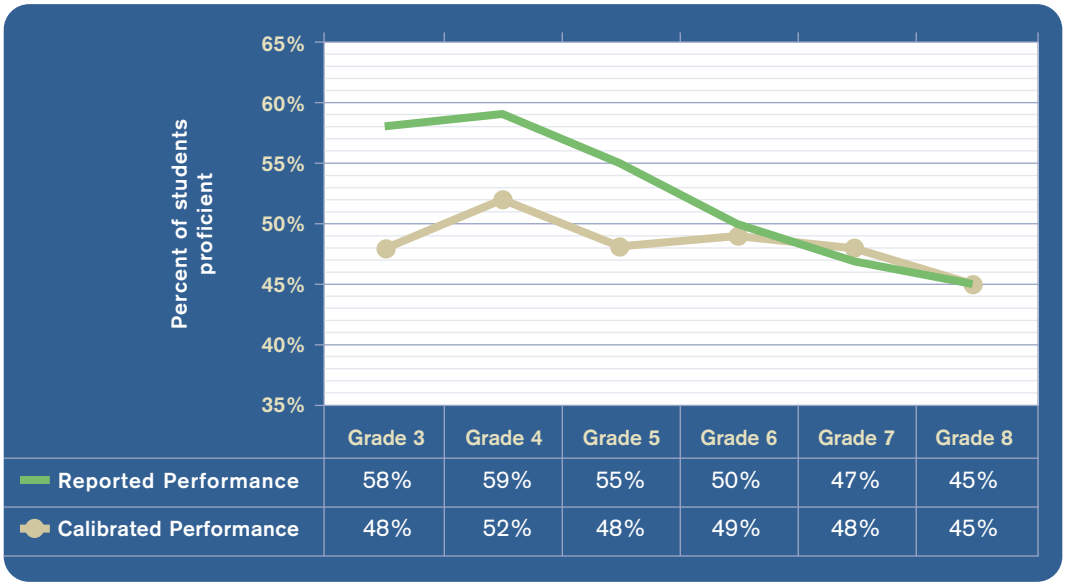
Figure 5 – Maine Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	65%	61%	58%	59%	60%	59%
Calibrated Performance	58%	60%	58%	61%	59%	59%

Note: This graphic shows, for example, that if Maine's grade-3 reading cut score was set at the same level of difficulty as its grade-8 cut score, 58 percent of third graders would achieve the proficient level, rather than 65 percent, as was reported by the state.

Figure 6 – Maine Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



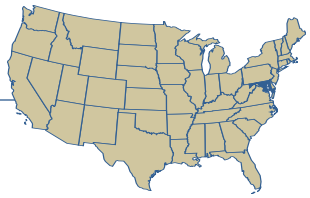
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	58%	59%	55%	50%	47%	45%
Calibrated Performance	48%	52%	48%	49%	48%	45%

Note: This graphic shows, for example, that if Maine's grade-3 mathematics cut score was set at the same level of difficulty as its grade-8 cut score, 48 percent of third graders would achieve the proficient level, rather than 58 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what students must know and be able to do in order to be considered proficient in reading and math, Maine is relatively high, at least compared with the other 25 states in this study. Maine's cut scores have been adjusted over the past several years, however, making them less challenging (although they are still more difficult than the majority of states in the current study). Also of note is the fact that Maine's proficiency cut scores in reading and math are not well calibrated across grades, particularly in math, where students who are proficient in third and fourth grade are not necessarily on track to be proficient by the eighth grade. Maine policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Maryland



Introduction

This study linked data from the 2005 and 2006 administrations of Maryland's reading test to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. (Mathematics data were not available because Maryland school districts only use the NWEA MAP tests in reading.) We found that Maryland's definition of proficiency in reading is somewhat lower than the median set by the other 25 states in this study. In other words, Maryland's reading tests are a bit below average in terms of difficulty.

In addition, the difficulty level of Maryland's reading tests decreased from 2005 to 2006 in some grades. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the Maryland test not being matched by learning gains on the Northwest Evaluation Association test. One striking finding of this study is that Maryland's reading cut scores are somewhat easier for elementary school students than for eighth-grade students (taking into account the differences in subject content and children's development). State policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Maryland School Assessment (MSA)

Maryland currently uses the Maryland School Assessment (MSA) which tests mathematics and reading in grades 3 to 8. The same sets of tests were used in spring 2005. The current study linked reading data from spring 2005 and spring 2006 MSA administrations to a common scale also administered in the 2005 and 2006 school years.

To determine the difficulty of Maryland's proficiency cut scores, we linked data from Maryland's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

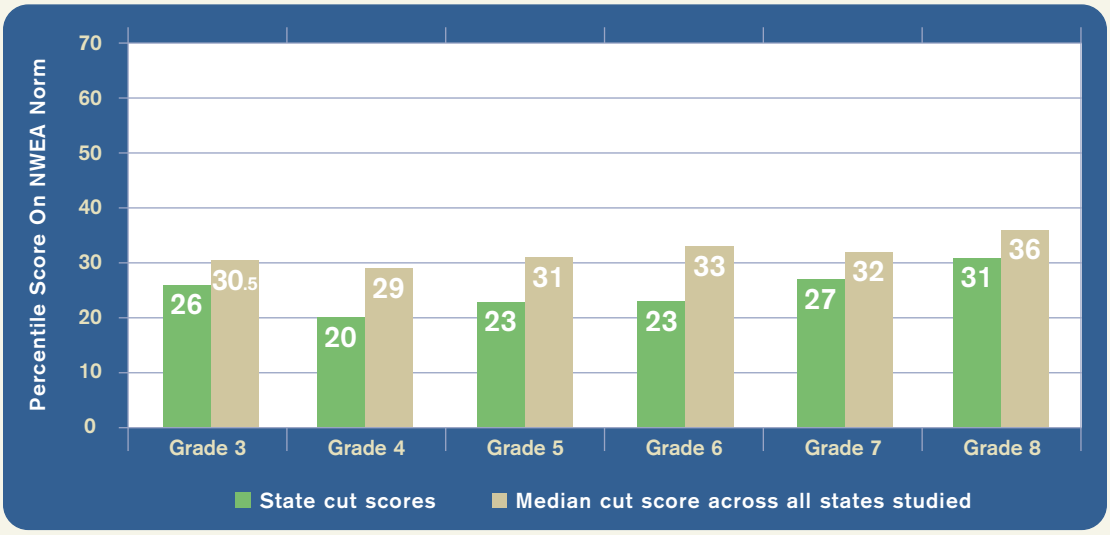
Part 1: How Difficult is Maryland's Definition of Proficiency in Reading?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of Maryland's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Maryland cut score on a test of equivalent difficulty. Figure 1 shows the difficulty of Maryland's **reading** proficiency cut scores in 2006 in relation to the median reading cut score for all the states in the study. Maryland's scores ranged between the 20th and 31st percentiles with respect to the NWEA norm group, with eighth grade being the most challenging.

Another way of assessing difficulty is to evaluate how Maryland's proficiency cut scores rank relative to other states. Table 1 shows that the Maryland cut scores generally rank in the lowest third in difficulty among the 26 states studied for this report.

Figure 1 – Maryland Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the cut scores of all 26 states reviewed in this study. Maryland's cut scores are consistently 4.5 to 10 percentile points below the median in grades 3 to 8.

Table 1 – Maryland Rank for Proficiency Cut Scores in Relation to 26 States, Reading, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	16	22	20	21	20	18

Note: This table ranks Maryland's reading cut scores relative to the cut scores of the other 25 states in the study, where 1 is highest and 26 is lowest.

Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Maryland's proficiency cut scores for the tests were mapped to their equivalent scores on NWEA's MAP assessment for the 2005 and 2006 school years. Cut score estimates for both years were possible for grades 3, 4, and 5.

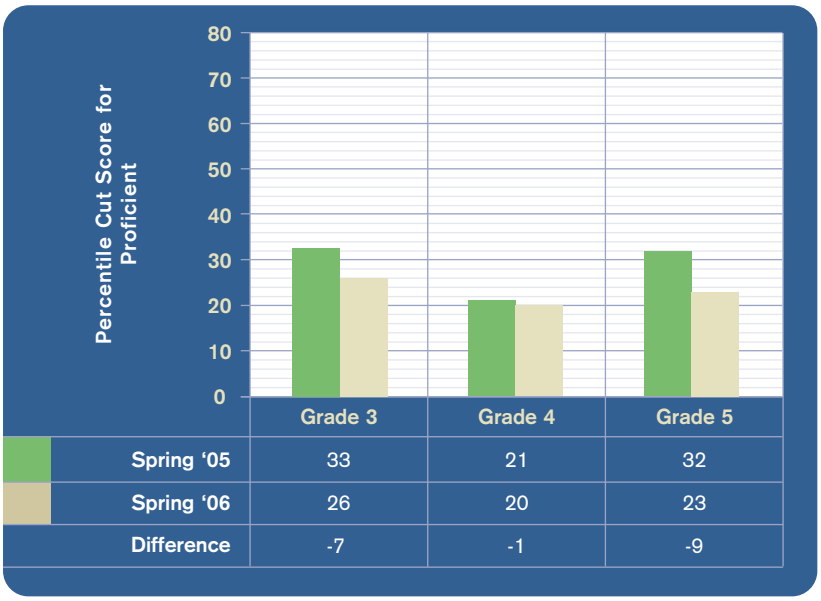
States may periodically re-adjust the cut scores they use to define proficiency in reading and mathematics, or update the exams used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Unintentional drift can occur even in states, such as Maryland, that maintained their proficiency levels.

Is it possible, then, to compare the proficiency scores between earlier administrations of Maryland's tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The MSA in 2005 and in 2006 can both be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to a meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the MSA in 2005 and 2006 on the MAP scale and ascertain whether the state test may have changed in difficulty.

In **reading**, Maryland's estimated cut scores decreased over this period in the third and fifth grade (see Figure 2), but there was essentially no change in the fourth-grade cut score. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the reading proficiency rate in 2006 to be 7 percent higher than in 2005 for third grade and 9 percent higher for fifth grade. (Maryland reported a 2 point gain for third graders and a 3 point gain for fifth graders over this period.)

Thus, one could fairly say that Maryland's third- and fifth-grade reading tests were easier to pass in 2006 than in 2005, while the fourth-grade test was about the same. As a result, improvements in the state's self-reported third- and fifth-grade proficiency rates during this period may not be entirely a product of improved achievement, while any improvements in the fourth-grade performance would signal real change in student performance.

Figure 2 – Estimated Differences in Maryland's Proficiency Cut Scores in Reading, 2005-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4	Grade 5
Spring '05	33	21	32
Spring '06	26	20	23
Difference	-7	-1	-9

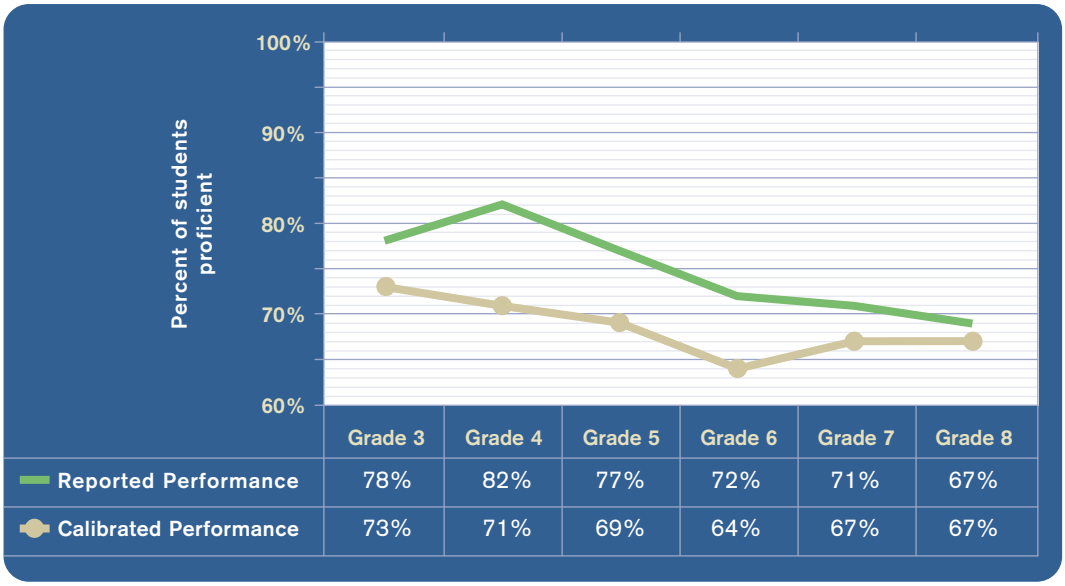
Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third-grade students in 2005 had to score at the 33rd percentile on the NWEA scale in order to be considered proficient, while a year later third graders had only to score at the 26th percentile to achieve proficiency. The changes in grade 4 were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Maryland's cut scores, we find that they are not well calibrated across grades. Figure 1 gave the relative difficulty of Maryland's 2006 reading cut scores across grades 3 to 8 (the "NCLB grades"), showing that cut scores in the upper grades tended to be more difficult than the cut scores in the lower grades. Figure 3 shows Maryland's reported reading performance on its state test compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades. This would lead to the conclusion that the higher rates of proficiency that the state has reported for students in lower grades are somewhat misleading, especially in grades 4, 5, and 6.

Figure 3 – Maryland Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



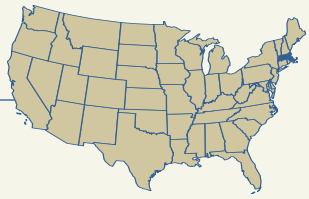
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	78%	82%	77%	72%	71%	67%
Calibrated Performance	73%	71%	69%	64%	67%	67%

Note: This graphic shows, for example, that if Maryland's grade-3 reading cut score were set at the same level of difficulty as its grade-8 cut score, 73 percent of third graders would achieve the proficient level, rather than 78 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what it takes for a student to be considered proficient in reading, Maryland is below the middle of the pack, at least compared with the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Maryland's standards to be at or just below the middle of the distribution of all states studied. From 2005 to 2006, Maryland's reading test became easier to pass, although not for all grades. As a result, Maryland's expectations are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. State policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Massachusetts



Introduction

This study linked data from the 2006 administration of Massachusetts's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Massachusetts's definitions of proficiency in reading and math are relatively high compared with the standards set by the other 25 states in the study, In other words, Massachusetts's tests are well above average in terms of difficulty.

However, unlike most of the states in this study, Massachusetts's proficiency cut scores for reading and English/language arts are less difficult in the later grades than in the earlier grades. Therefore, reported results for younger students may underestimate the number who are on track to be proficient in eighth-grade reading. Massachusetts policymakers might consider adjusting their reading cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Massachusetts Comprehensive Assessment System (MCAS)

Massachusetts currently uses the Massachusetts Comprehensive Assessment System (MCAS), which tests mathematics and reading/ELA in grades 3 to 8 and grade 10, and high school science and technology in grades 9 and 10. The current study linked reading and math data from spring 2006 MCAS administrations to a common scale also administered in the 2006 school year.

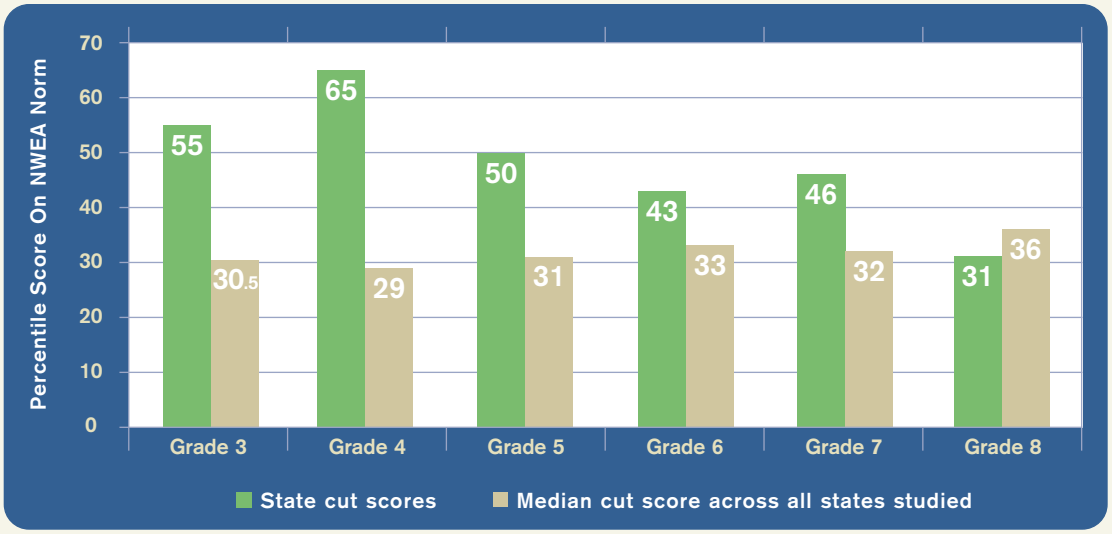
To determine the difficulty of Massachusetts's proficiency cut scores, we linked data from Massachusetts's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (For more details on how this was done, please see the methodology section of this report.)

Part 1: How Difficult are Massachusetts's Definitions of Proficiency in Reading and Math?

One way to assess the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to leap? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? We know because only one (or perhaps none) of those same 100 individuals would successfully meet that level of challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

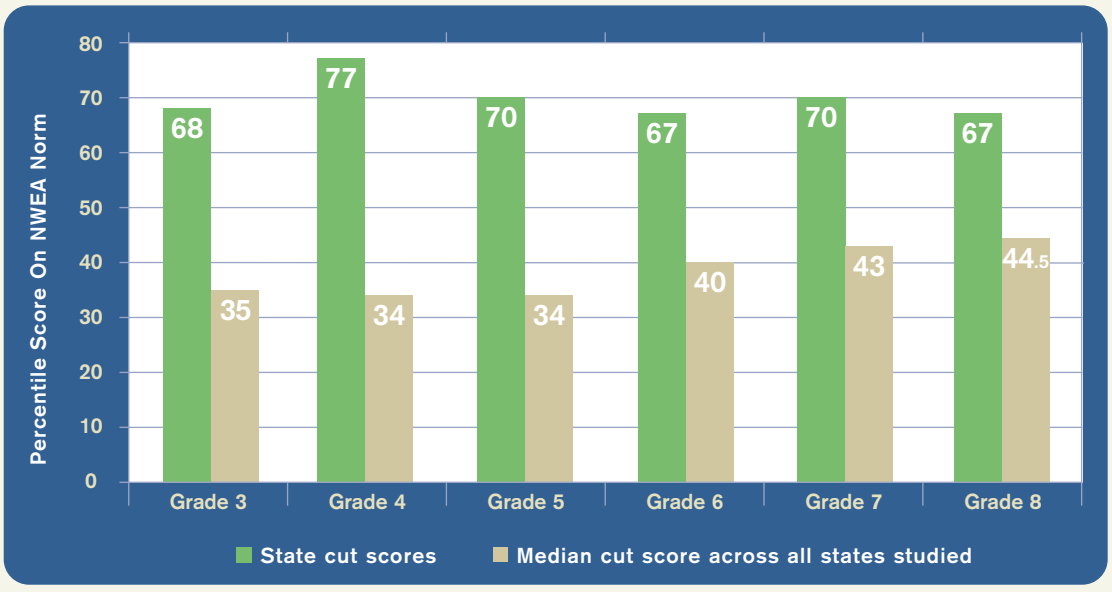
Applying the concept to this task, we evaluated the difficulty of the Massachusetts proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the cut score on a test of equivalent difficulty. The following two figures show the difficulty of Massachusetts's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the other states in the study, and compared with the NWEA norm group. The proficiency cut scores for **reading** in Massachusetts ranged between the 31st and 65th percentiles in the norm group, with the fourth-grade cut score being most challenging. In **mathematics**, the cut scores ranged between the 67th and 77th percentiles with fourth grade again being most challenging.

Figure 1 – Massachusetts Reading Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of all 26 states reviewed in this study. Massachusetts is consistently above average—as much as 36 percentile points above the median in fourth grade—except for eighth grade, when it falls 5 percentiles below the median.

Figure 2 – Massachusetts Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: Massachusetts math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of all 26 states reviewed in this study. The math cut scores are consistently 22.5 to 43 percentile points above the median.

Massachusetts's reading cut scores are consistently above the median difficulty of the 26 states that we examined, except in grade 8. Massachusetts's mathematics cut scores are above the median in every grade. Note, too, that the reading cut scores are consistently less difficult than the corresponding mathematics cut scores. Thus, reported differences in achievement on the MCAS between reading and mathematics might be more a product of differences in cut scores than in actual student achievement. In other words, Massachusetts students may be performing worse in reading or better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

Another way of assessing difficulty is to evaluate how Massachusetts's proficiency cut scores rank relative to other states. Table 1 shows that the Massachusetts cut scores rank at the very top in difficulty among the 26 states in this study, except in eighth grade reading.

Table 1 – Massachusetts Reading and Mathematics Cut Scores for Proficient Performance, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	2	1	4	4	4	18
Mathematics	2	1	2	1	1	2

Note: This table ranks Massachusetts's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

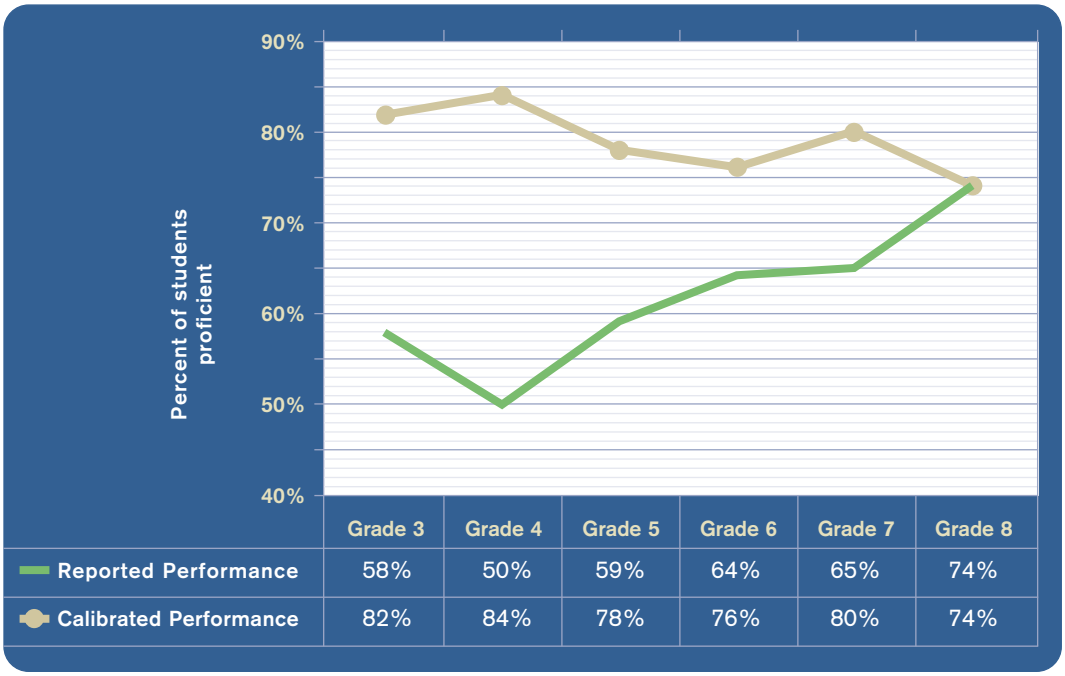
Part 2: Calibration across Grades*

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Massachusetts's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 illustrated that Massachusetts's reading and mathematics proficiency cut scores differed across grades in terms of their relative difficulty. These figures showed that the reading cut scores at the earlier grades were somewhat more difficult than the cut scores at the later grades. (The opposite is true in most states studied.) The mathematics cut scores, however, were fairly consistent across grades. These differing patterns are reflected in Figures 3 and 4, which show Massachusetts's reported performance in reading and mathematics on the state tests, and how those proficiency rates would look if the cut scores were all calibrated to the grade-8 standard. In Figure 3, we see that the state-reported proficiency rates underestimate the proportion of students who are on track to eventually meet the easier eighth-grade reading requirements. In Figure 4, we see less difference between the calibrated and actual reported proficiency rates, since the math cut scores themselves are much more consistent across grades.

* Massachusetts was one of seven states in this study for which cut score estimates could be determined only for one year. Therefore, it was not possible to examine whether its cut scores have changed over time.

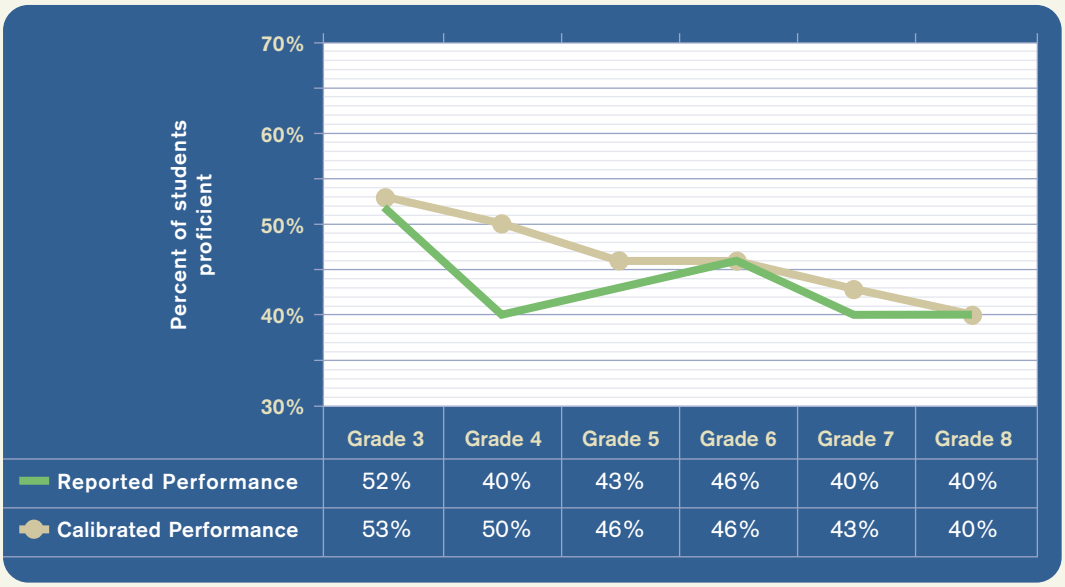
Figure 3 – Massachusetts Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	58%	50%	59%	64%	65%	74%
Calibrated Performance	82%	84%	78%	76%	80%	74%

Note: This graphic means that, for example, if Massachusetts's grade-3 reading cut score were set at the same level of difficulty as its grade-8 cut score, 82 percent of third graders would achieve the proficient level, rather than 58 percent, as was reported by the state.

Figure 4 – Massachusetts Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	52%	40%	43%	46%	40%	40%
Calibrated Performance	53%	50%	46%	46%	43%	40%

Note: This graphic shows, for example, that if Massachusetts's grade-4 mathematics cut score were set at the same level of difficulty as its grade-8 cut score, 50 percent of fourth graders would achieve the proficient level, rather than 40 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what it takes for a student to be considered proficient in reading and math, Massachusetts is relatively high, compared with the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales,* which also found Massachusetts's standards to be in the top third among all states studied. However, Massachusetts's grade-8 reading cut score is significantly less difficult than in earlier grades. State policymakers might consider adjusting their reading standards across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

Michigan

Introduction

This study linked data from the 2003 and 2005 administrations of Michigan's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Michigan's definitions for proficiency in reading and mathematics are less difficult than the standards set by most of the other 25 other states in this study. In other words, Michigan's tests are well below average in terms of difficulty.

In addition, the level of difficulty of Michigan's tests decreased somewhat from 2003 to 2005—the No Child Left Behind era—although not in all grades. One finding of this study is that Michigan's standards are dramatically lower for third-grade students than for eighth-grade pupils (taking into account the differences in subject content and children's development). State policymakers might consider adjusting the standards to ensure equivalent difficulty at all grades so that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Michigan Educational Assessment Program (MEAP)

Michigan currently uses a fall assessment called the Michigan Educational Assessment Program (MEAP), which tests English/language arts and mathematics in grades 3 through 8, science in grades 5 and 8, and social studies in grades 6 and 9. The current study linked data from fall 2003 and fall 2005 administrations to a common scale also administered in the 2003 and 2005 school years. To determine the difficulty of Michigan's proficiency cut scores, we linked data from Michigan's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered "proficient.") This was done by analyzing the reading and math results of a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Michigan's Definitions of Proficiency in Reading and Math?

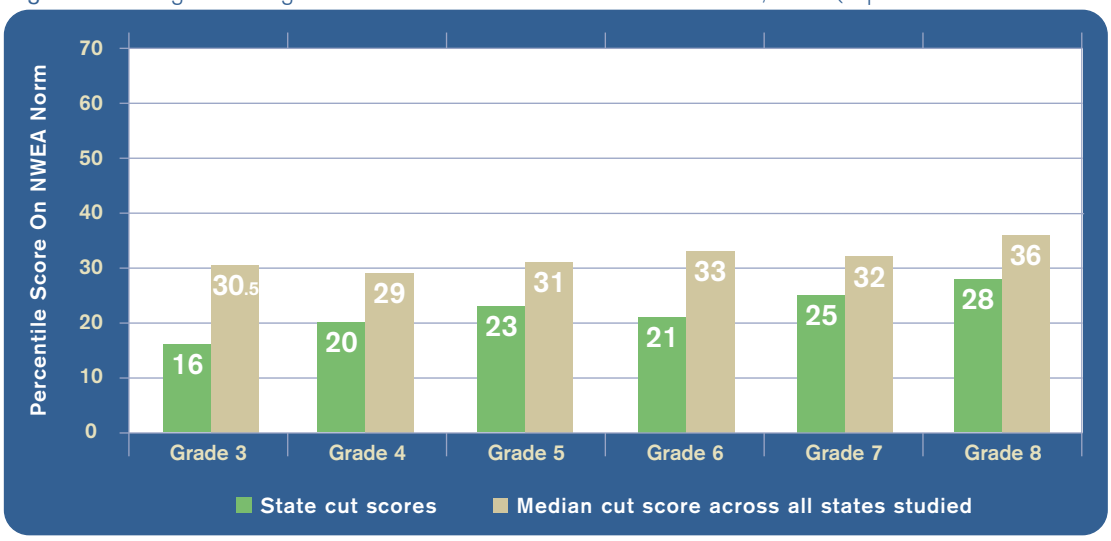
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of Michigan's proficiency standards by estimating the proportion of students in NWEA's national norm group who would perform above the Michigan standard on a test of equivalent difficulty. The following two figures show the difficulty of Michigan's proficiency standards for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median of all the states in the study. The proficiency cut scores for **reading** in Michigan ranged between the 16th and 28th percentiles for the norm group, with the eighth-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 6th and 35th percentiles, with seventh grade being most challenging.

Figures 1 and 2 show us that Michigan's cut scores in both reading and mathematics are consistently less difficult than the median standards of the other states in the study and well below the capabilities of the average student within the NWEA norm group.

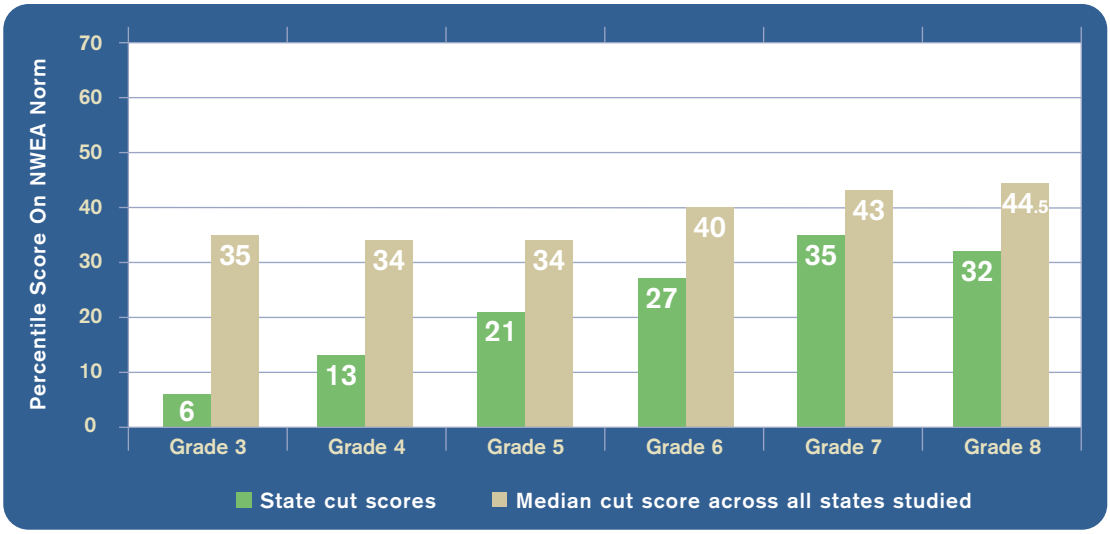
Another way of assessing difficulty is to evaluate how Michigan's proficiency cut scores rank relative to other 25 states within the study. Table 1 shows that the Michigan standards generally rank among the lowest in terms of difficulty.

Figure 1 – Michigan Reading Cut Scores in Relation to All 26 States Studied, 2005 (Expressed in MAP Percentile)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of all 26 states reviewed in this study. Michigan's reading cut scores are consistently 7 to 14.5 percentiles below the median.

Figure 2 – Michigan Mathematics Cut Scores in Relation to All 26 States Studied, 2005
(Expressed in MAP Percentile)



Note: Michigan's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of all 26 states reviewed in this study. Michigan's cut scores are consistently below the median, particularly in the early years, when the math cut score is as much as 29 percentiles below the median.

Table 1 – Michigan Reading and Mathematics Standards for Proficient Performance, 2005

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	21	22	20	22	21	20
Mathematics	24	24	23	21	21	19

Note: This table ranks Michigan's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

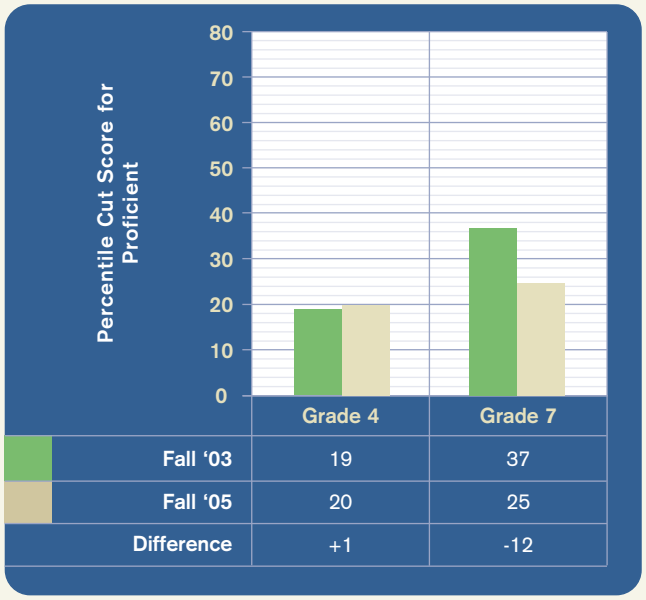
Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Michigan's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2003 and 2005 school years. Cut score estimates for both years were available for grades four and seven in reading, and for grades four and eight in mathematics.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the tests used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. In Michigan's case, the state adopted a new scale and new cut scores effective for the fall 2005 testing season.

Is it possible, then, to compare the proficiency scores between earlier administrations of Michigan tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. MEAP in 2003 and MEAP in 2005 can both be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the MEAP in 2003 and 2005 on the MAP scale and ascertain whether the test may have changed in difficulty.

Figure 3 – Estimated Difference in Michigan's Proficiency Cut Scores in Reading, 2003-2005.



	Grade 4	Grade 7
Fall '03	19	37
Fall '05	20	25
Difference	+1	-12

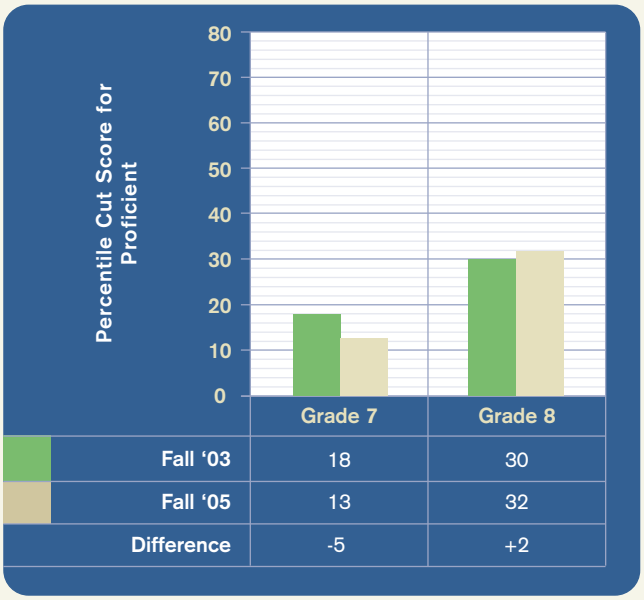
Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, seventh-grade students in 2003 had to score at the 37th percentile of the NWEA norm in order to be considered proficient, while in 2005 seventh graders had only to score at the 25th percentile to achieve proficiency. The change in grade 4 was within the margin of error (in other words, too small to be considered substantive).

In **reading**, there was no substantive change in the estimated fourth-grade standard over the two-year period, but a large decrease in the seventh-grade standard (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the seventh-grade reading proficiency rate in 2005 to rise by about 12 percent over the 2003 level simply because of the easier standard. (Michigan reported a 15-point gain for seventh graders over this period.)

Michigan's estimated **mathematics** cut scores showed the reverse pattern, with a moderate decrease in the fourth-grade standard and essentially no change in the eighth-grade standard (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, the less difficult fourth-grade standard in 2005 would elicit a proficiency rating that was five percent higher than the 2003 level. (Michigan reported a 17-point gain for fourth graders over this period.)

Thus, one could fairly say that Michigan's seventh-grade reading and fourth-grade math tests were easier to pass in 2005 than in 2003, but the tests in the other observed grades remained about the same. As a result, state-reported gains in fourth-grade math and seventh-grade reading proficiency rates during this period may not be entirely a product of improved achievement.

Figure 4 – Estimated Differences in Michigan's Proficiency Cut Scores in Mathematics, 2003-2005 (Expressed in MAP Percentiles)



	Grade 7	Grade 8
Fall '03	18	30
Fall '05	13	32
Difference	-5	+2

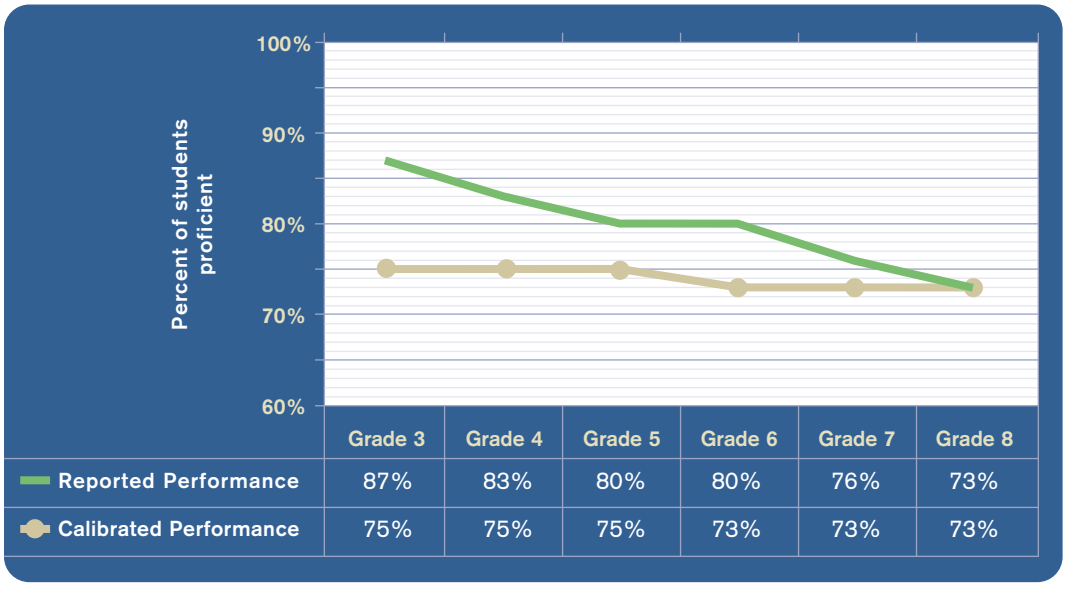
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, fourth-grade students in 2003 had to score at the 18th percentile nationally in order to be considered proficient, while in 2005, fourth graders only had to score at the 13th percentile to achieve proficiency. The change in grade 8 was within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Michigan's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 above showed that Michigan's upper-grade cut scores in reading and mathematics were generally more challenging than the standards in the lower grades. The two figures that follow show Michigan's reported performance on its state test in reading (Figure 5) and mathematics (Figure 6) compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the standard are removed, student performance is much more consistent across grades. This would lead to the conclusion that the higher rates of proficiency that the state has reported for lower grades students are somewhat misleading.

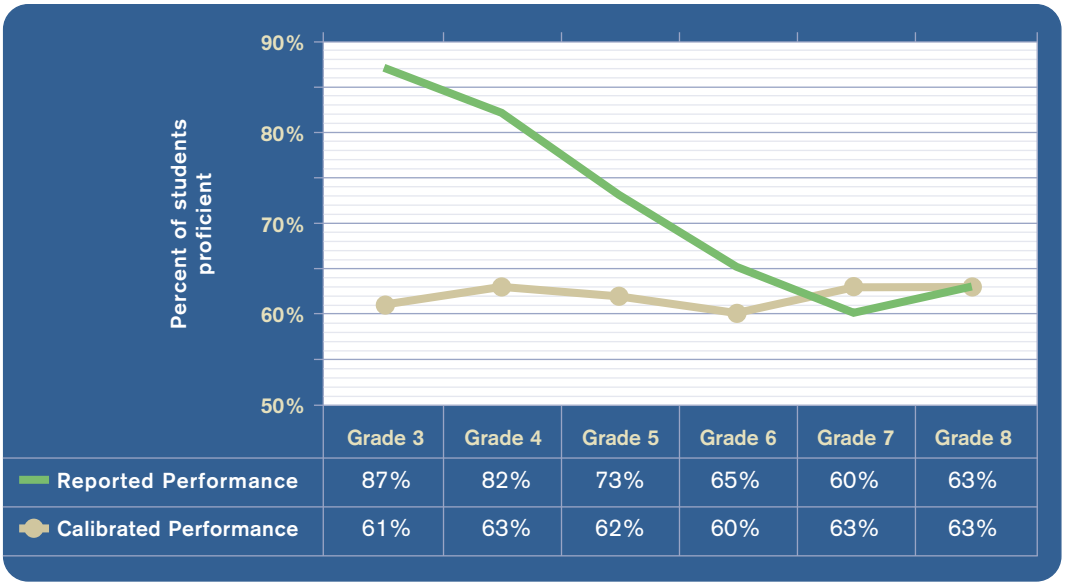
Figure 5 – Michigan Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	87%	83%	80%	80%	76%	73%
Calibrated Performance	75%	75%	75%	73%	73%	73%

Note: This graphic shows, for example, that if Michigan's grade-3 reading standard were set at the same level of difficulty as its grade-8 standard, 75 percent of third graders would achieve the proficient level, rather than 87 percent, as was reported by the state.

Figure 6 – Michigan Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



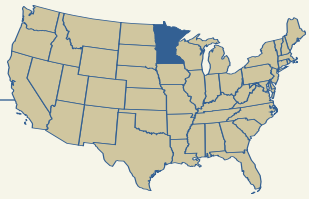
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	87%	82%	73%	65%	60%	63%
Calibrated Performance	61%	63%	62%	60%	63%	63%

Note: This graphic shows, for example, that if Michigan's grade-3 mathematics standard were set at the same level of difficulty as its grade-8 standard, 61 percent of third graders would achieve the proficient level, rather than 87 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what it takes for a student to be considered proficient in reading and math, Michigan is low compared to the other 25 states in this study. (This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Michigan standards to be in the bottom half or bottom third of the distribution of all states studied for mathematics.) From 2003 to 2005, its reading and mathematics proficiency standards have declined somewhat, though not for all grades. In addition, Michigan's expectations are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. Michigan policymakers might consider adjusting their standards across the board but especially in the earlier grades, so that parents and schools can be assured that young students scoring at the proficient level are truly prepared for success later in their educational careers.

Minnesota



Introduction

This study linked data from the 2003 and 2006 administrations of Minnesota's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Minnesota's definitions of proficiency in reading and mathematics are somewhat more difficult than the standards set by many of the other 25 states in this study. In other words, Minnesota's tests are above average in terms of difficulty.

The level of difficulty changed some from 2003 to 2006—the No Child Left Behind era—although the direction of that change has varied by grade level. Minnesota's current test appears to be easier in third grade and harder in eighth grade than the test it replaced. As a result, Minnesota's cut scores are now dramatically lower for third-grade students than for eighth-grade pupils (taking into account the differences in subject content and children's development). Minnesota policymakers might consider adjusting the cut scores to ensure equivalent difficulty at all grades so that elementary school students are on track to be proficient in the later grades.

What We Studied: Minnesota's Assessment Program

The Minnesota Comprehensive Assessment II (MCA-II) is currently used for students in grades 3 through 8. The MCA-II is referred to as a standards-referenced test, which means that its primary purpose is to assess how students perform relative to expectations for the grades in which they are enrolled. MCA-II replaced the Minnesota Comprehensive Assessment I, which was administered in grades 3 and 5 until 2005. Prior to 2005, the Minnesota Basic Skills Test (BST) was administered to students in grade 8.

The MCA-II is designed to align with Minnesota's standards and benchmarks for each grade level.

To determine the difficulty of Minnesota's proficiency cut scores, we linked reading and math data from state tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state assessment and the NWEA test. (The methodology section of this report explains how performance was compared.)

Part 1: How Difficult are Minnesota's Definitions of Proficiency in Reading and Math?

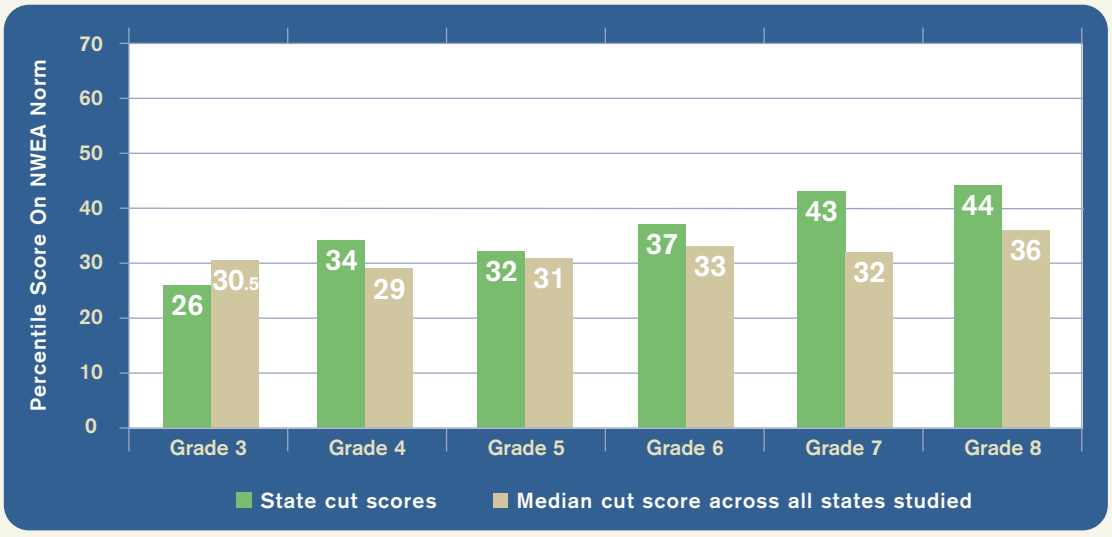
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of Minnesota's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Minnesota cut score on a test of equivalent difficulty. The following two figures show the difficulty of Minnesota's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Minnesota ranged between the 26th and 44th percentiles for the norm group, with the eighth-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 30th and 54th percentiles with fifth grade being most challenging.

Except in grade 3, Minnesota's cut scores in both reading and math are above the median difficulty among the states studied. Note, though, that Minnesota's cut scores for reading are lower than those for mathematics. (This was the case for the majority of states studied.) Thus, reported differences in achievement on the MCA-II between reading and mathematics might be more a product of differences in cut scores than in actual student achievement. In other words, Minnesota students may be performing worse in reading or better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

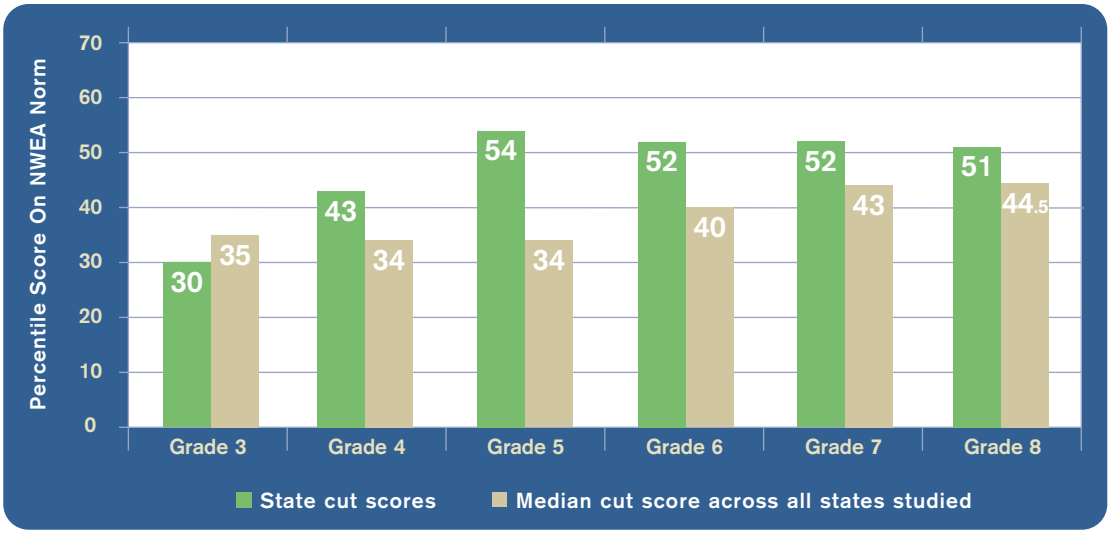
Another way of assessing difficulty is to evaluate how Minnesota's proficiency cut scores rank relative to other states. Table 1 shows that the Minnesota cut scores generally rank in the upper half in difficulty among the 26 states studied for this report. Its reading cut scores in grade 7 and mathematics cut scores in grade 5 rank among the top four to five states in difficulty.

Figure 1 – Minnesota Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of all 26 states reviewed in this study. Except for grade 3, Minnesota's reading cut scores are all above the median.

Figure 2 – Minnesota Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: Minnesota's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of all 26 states reviewed in this study. Except in grade 3, Minnesota's cut scores are consistently 6.5 to 20 percentile points above the median.

Table 1 – Minnesota Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	16	6	11	10	5	6
Mathematics	14	8	4	6	7	10

Note: This table ranks Minnesota's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

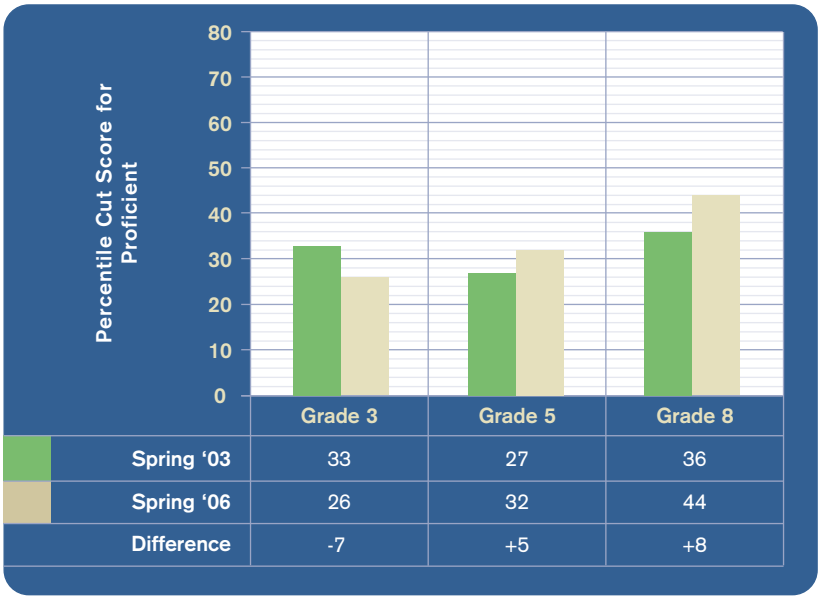
Part 2: Changes in Cut Scores over Time

In order to measure their consistency, Minnesota's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2003 and 2006 school years. Because in 2003 the Minnesota Comprehensive Assessment (called the MCA-I) was administered only in grades 3 and 5 and the BST was given only in grade 8, the estimates of change over time are limited to these grades.

After changing over from the MCA-I and BST to MCA-II, the Minnesota Department of Education established new cut scores for all grades. Because the tests were different in various ways, changes in the definition of proficiency were to be expected. For that reason, the Minnesota Department of Education cautions that results from the MCA-I and BST should not be considered equivalent to the results from the MCA-II series of exams.

Is it possible anyway to compare the proficiency scores between earlier administrations of Minnesota tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. Although the MCA-I, MCA-II, and BST's are different measures, they can all be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the Minnesota tests in 2003 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty.

Figure 3 – Estimated Differences in Minnesota's Proficiency Cut Scores in Reading, 2003-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 5	Grade 8
Spring '03	33	27	36
Spring '06	26	32	44
Difference	-7	+5	+8

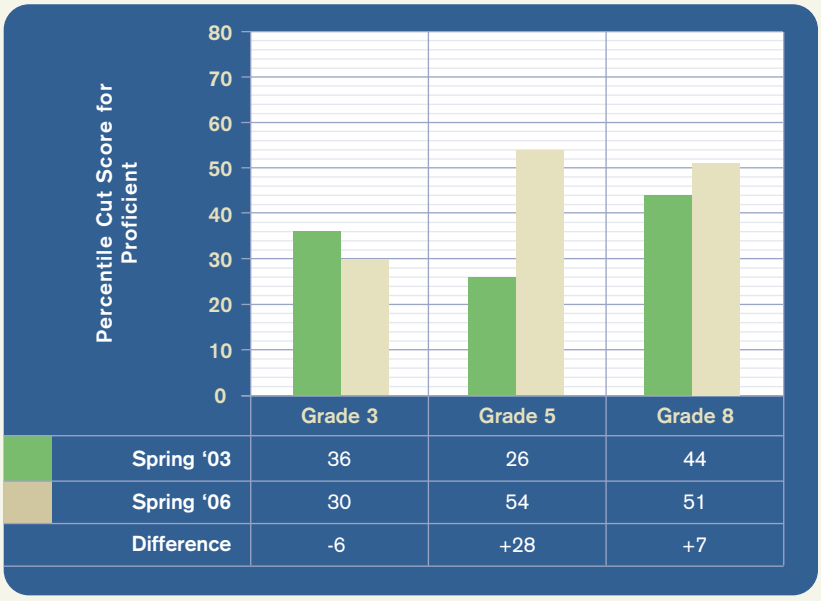
Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third-grade students in 2003 had to score at the 33rd percentile on the NWEA norm in order to be considered proficient, while in 2006 third graders only had to score at the 26th percentile to achieve proficiency. The change in grade 5 was within the margin of error (in other words, too small to be considered substantive).

In **reading**, Minnesota's estimated cut scores decreased over this three-year period in the third grade (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the third-grade reading proficiency rate in 2006 to be 7 percent higher than in 2003. (Minnesota reported a 5-point gain for third graders over this period.) For grade 8, the reading proficiency cut score rose. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the eighth-grade reading proficiency rate to decline by 8 percent. (Minnesota reported a 17-point decline for eighth graders over this period.)

In **mathematics**, Minnesota showed increases in estimates of their fifth- and eighth-grade mathematics cut scores (see Figure 4). These were large enough to cause a 28 percent drop in the expected proficiency rating for fifth grade, and a 7 percent drop in the pass rate for eighth grade. (Minnesota reported an 18-point decline for fifth graders and a 15-point decline for eighth graders over this period.)

Thus, one could fairly say that Minnesota's third-grade test in reading was easier to pass in 2006 than in 2003, while the eighth-grade reading and the fifth- and eighth-grade math tests became substantively harder to pass. As a result, improvements in the state-reported third grade proficiency rate during this period may not be entirely a product of improved achievement, while real improvements in other areas may be masked somewhat by the increased difficulty of the state's proficiency cut scores at these grades.

Figure 4 – Estimated Differences in Minnesota's Proficiency Cut Scores in Mathematics, 2003-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 5	Grade 8
Spring '03	36	26	44
Spring '06	30	54	51
Difference	-6	+28	+7

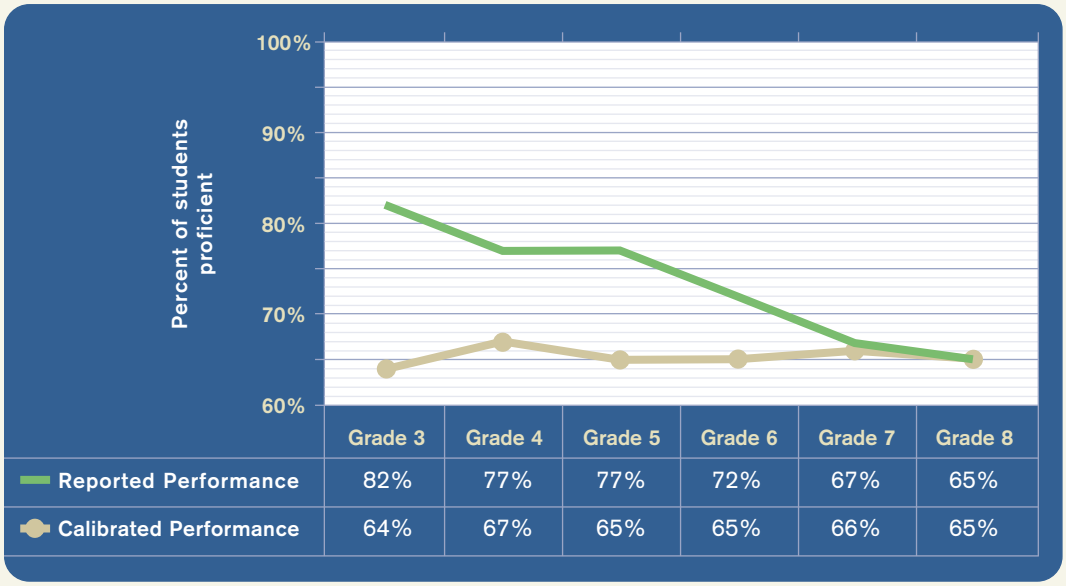
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, fifth-grade students in 2003 had to score at the 26th percentile on the NWEA norm in order to be considered proficient, while by 2006 fifth graders had to score at the 54th percentile to achieve proficiency. The change in grade 3 was within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Minnesota's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed that, as in most other states in this study, Minnesota's upper-grade cut scores in reading and math in 2006 were considerably more challenging than the cut scores in the lower grades, particularly grade 3. The two figures that follow show Minnesota's reported performance in reading (Figure 5) and mathematics (Figure 6) on its state test and the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut scores are taken into account, student performance is more consistent across grades. This would lead to the conclusion that the higher proficiency rates reported by the state for students in earlier grades are somewhat misleading.

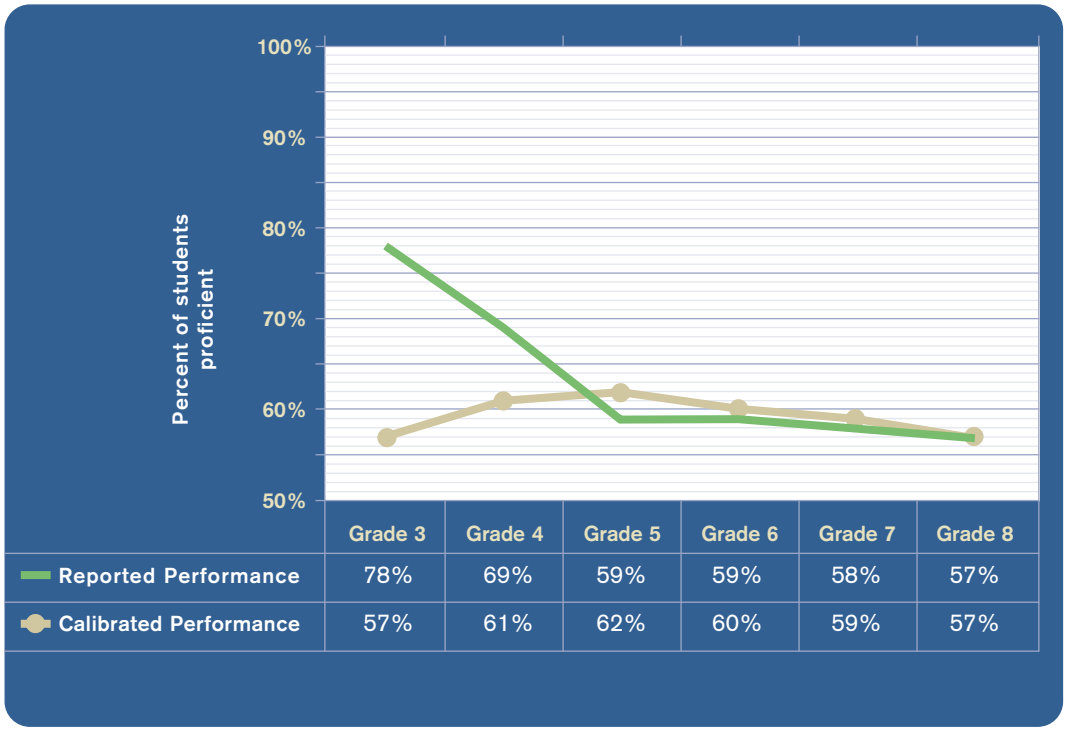
Figure 5 – Minnesota Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	82%	77%	77%	72%	67%	65%
Calibrated Performance	64%	67%	65%	65%	66%	65%

Note: This graphic shows, for example, that if Minnesota's grade-3 reading cut score were set at the same level of difficulty as its grade-8 cut score, only 64 percent of third graders would achieve the proficient level, rather than 82 percent, as reported by the state.

Figure 6 – Minnesota Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



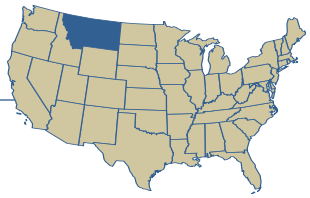
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	78%	69%	59%	59%	58%	57%
Calibrated Performance	57%	61%	62%	60%	59%	57%

Note: This graphic shows that, for example, if Minnesota's grade-3 mathematics cut score were set at the same level of difficulty as its grade-8 cut score, only 57 percent of third graders would achieve the proficient level, rather than 78 percent, as was reported by the state.

Policy Implications

When setting the cut scores for what it takes for a student to be considered proficient in reading and math, Minnesota is relatively high, at least compared with the other 25 states in this study. In recent years, the state has adjusted the difficulty of these cut scores—making them more challenging in the later grades and less so in the early ones. As a result, Minnesota's expectations are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. State policymakers might consider adjusting their standards across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

Montana



Introduction

This study linked data from the 2004 and 2006 administrations of Montana's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Montana's definitions of proficiency are relatively consistent with the standards set by the other 25 states in the study with respect to reading, but relatively difficult compared with other states with respect to mathematics. In other words, Montana's reading tests are about average and its math tests are harder than average.

The level of difficulty changed some from 2004 to 2006—the No Child Left Behind era. Montana's reading tests became easier at both the fourth- and eighth-grade levels, while its math test became easier in fourth grade and much harder in eighth grade. There are many possible explanations for these declines in our estimates of Montana's cut scores (see pp. 34-35 of the main report), which were caused by learning gains on the state test not being matched by learning gains on the Northwest Evaluation Association test. As a result, Montana's cut scores are less difficult in the early grades than they are for eighth-grade pupils, especially in mathematics (taking into account the differences in subject content and children's development). Montana policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

What We Studied: Montana Criterion-Referenced Test (Montana CRT)

Montana currently uses an assessment called the Montana Criterion-Referenced Test (Montana CRT) which tests mathematics and reading in grades 3 through 8 and grade 10. The same sets of tests were used in spring 2004 to test students in mathematics and reading in grades 4, 8, and 10. The current study linked data from spring 2004 and spring 2006 administrations to a common scale also administered in the 2004 and 2006 school years.

To determine the difficulty of Montana's proficiency cut scores, we linked reading and math data from Montana's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Montana's Definitions of Proficiency in Reading and Math?

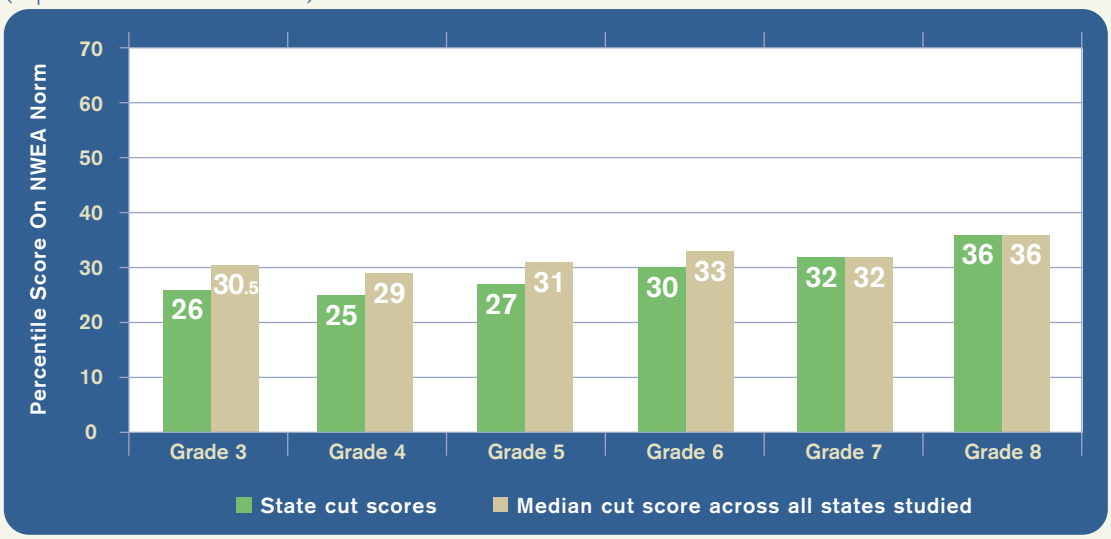
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of Montana's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Montana cut score on a test of equivalent difficulty. The following two figures show the difficulty of Montana's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all states in the study. The proficiency cut scores for **reading** in Montana ranged between the 25th and 36th percentiles for the norm group, with the eighth-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 40th and 60th percentiles, with eighth grade again being most challenging.

In most grades, Montana's cut scores for reading proficiency are close to the median level of difficulty, compared with the other states in the study. For mathematics, however, Montana's proficiency cut scores are generally above the median. Note, also, that Montana's cut scores for reading are relatively lower than for math. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Montana students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentages of pupils passing state tests in those subjects.

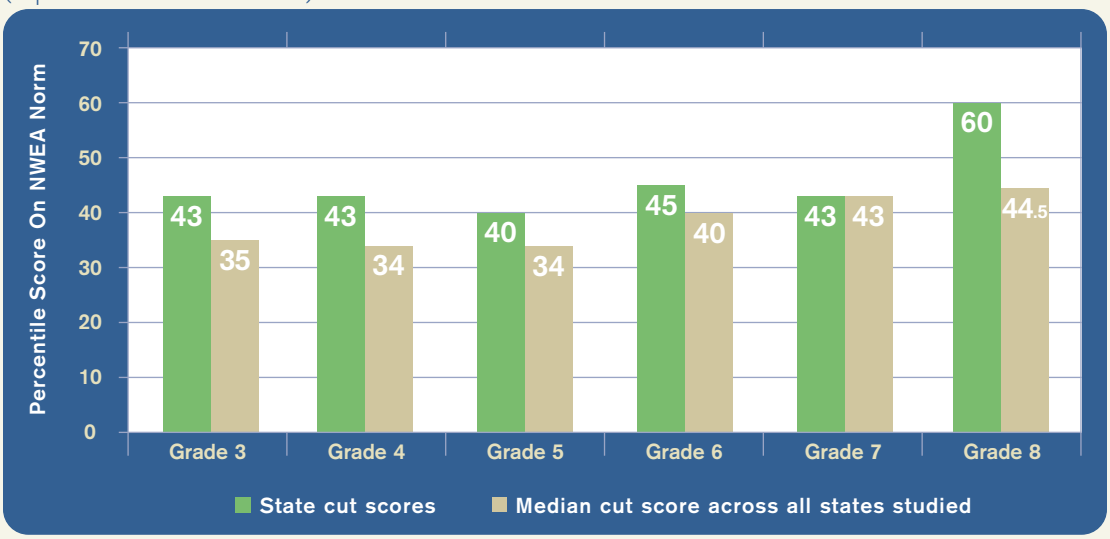
Another way of assessing difficulty is to evaluate how Montana's proficiency cut scores rank relative to other states. Table 1 shows that the Montana reading cut scores generally rank in the lower half in difficulty among the 26 states studied, and the upper half for mathematics. Its eighth-grade math cut score ranks among the top three across all states studied.

Figure 1 – Montana Reading Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut scores of all 26 states reviewed in this study. Montana's cut scores are slightly below the median except in seventh and eighth grades where the state's cut scores are at the median.

Figure 2 – Montana Mathematics Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: Montana's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of all 26 states reviewed in this study. Montana's cut scores are consistently 5 to 15.5 percentile points above the median except for seventh grade, which is at the median.

Table 1 – Montana Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	16	17	17	17	13	9
Mathematics	6	8	10	8	12	3

Note: This table ranks Montana's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

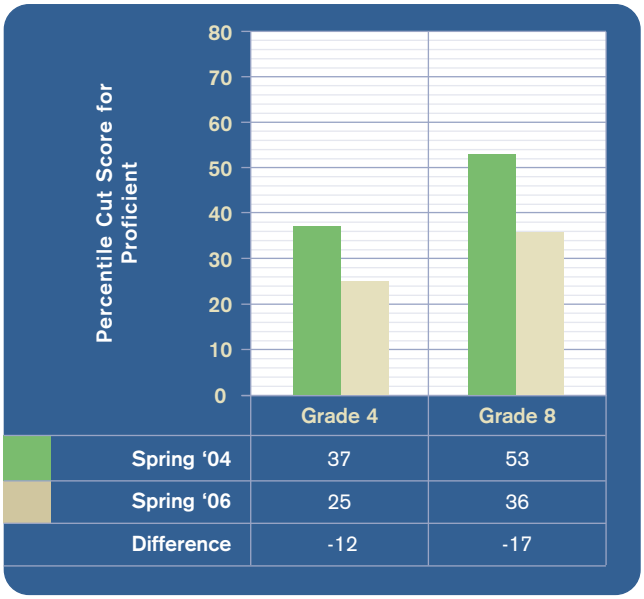
Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Montana's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2004 and 2006 school years. Information about proficiency cut scores for both school years was available for grades 4 and 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the exams used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Unintentional drift can occur even in states, such as Montana, that maintained their proficiency levels.

Is it possible, then, to compare the proficiency scores between earlier administrations of Montana tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The Montana CRT in 2004 and Montana CRT in 2006 can both be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the CRT in 2004 and 2006 on the MAP scale and ascertain whether the state test may have changed in difficulty.

Figure 3 – Estimated Differences in Montana's Proficiency Cut Scores in Reading, 2004-2006 (Expressed in MAP Percentiles)



	Grade 4	Grade 8
Spring '04	37	53
Spring '06	25	36
Difference	-12	-17

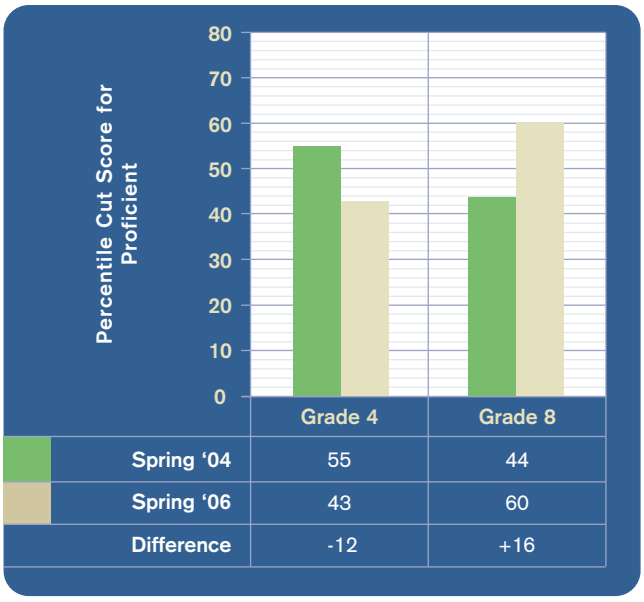
Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, fourth-grade students in 2004 had to score at the 37th percentile on the NWEA norm in order to be considered proficient, while in 2006 fourth graders had only to score at the 25th percentile to achieve proficiency.

Montana's estimated **reading** cut scores show large decreases for fourth and eighth grades over this two-year period (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the reading proficiency rate in 2006 to be 12 percent higher than in 2004 for grade 4, and 17 percent higher for grade 8. (Montana reported a 14-point gain for fourth graders and an 18-point gain for eighth graders over this period.)

Montana's estimated **mathematics** cut scores also show a decrease in the difficulty for fourth grade (Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, this would likely yield an increased proficiency rate of 12 percent. The eighth-grade cut scores increased dramatically, however, enough to cause a 16 percent drop in the expected proficiency rating for eighth grade. (Montana reported a 19-point gain for fourth graders and a 7-point decline for eighth graders over this period.)

Thus, one could fairly say that Montana's fourth-grade tests in both reading and mathematics were easier to pass in 2006 than in 2004, while the eighth-grade tests were easier in reading and harder in math. As a result, some apparent improvements in state-reported fourth-grade proficiency rates during this period may not be entirely a product of improved achievement, while any improvements in eighth-grade mathematics performance may be masked by the more difficult proficiency cut score.

Figure 4 – Estimated Differences in Montana's Proficiency Cut Scores in Mathematics, 2004-2006 (Expressed in MAP Percentiles)



	Grade 4	Grade 8
Spring '04	55	44
Spring '06	43	60
Difference	-12	+16

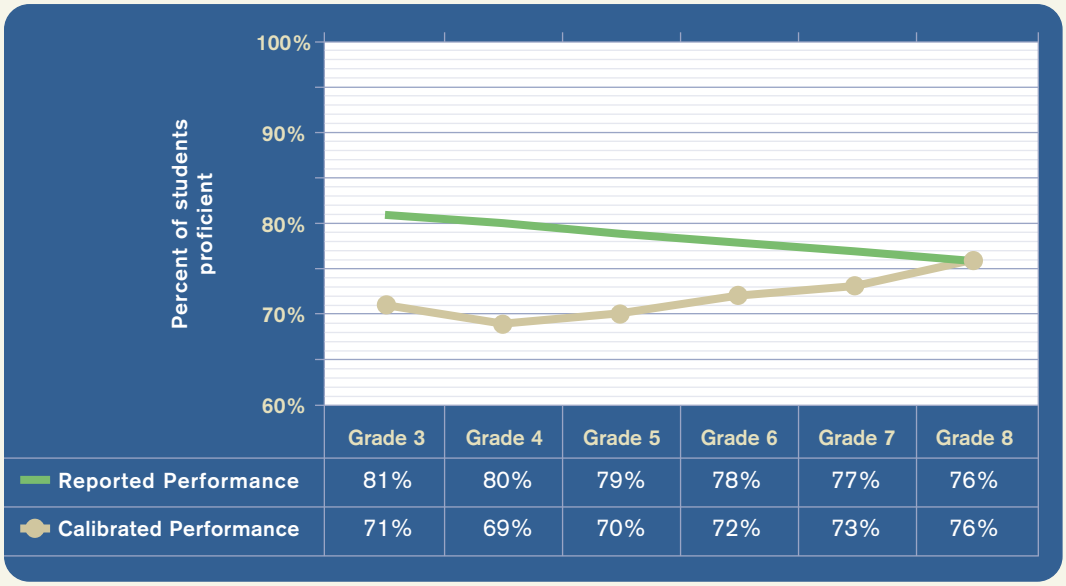
Note: This graphic shows how the degree of difficulty in achieving proficiency in math has changed. For example, fourth-grade students in 2004 had to score at the 55th percentile on the NWEA norm in order to be considered proficient, while in 2006 fourth graders only had to score at the 43rd percentile to achieve proficiency.

Part 3: Calibration across Grades

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth-grade cut score is no more or less difficult for eighth graders to achieve than the third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 gave the relative difficulties of the reading and mathematics cut scores across grades, showing that the upper-grade cut scores in reading and mathematics were more difficult than those in the lower grades. The following two figures show Montana's reported performance in reading (Figure 5) and mathematics (Figure 6) on the state test and the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance at the lower grades is less likely to overestimate the percentage of students on track to meet eighth-grade expectations.

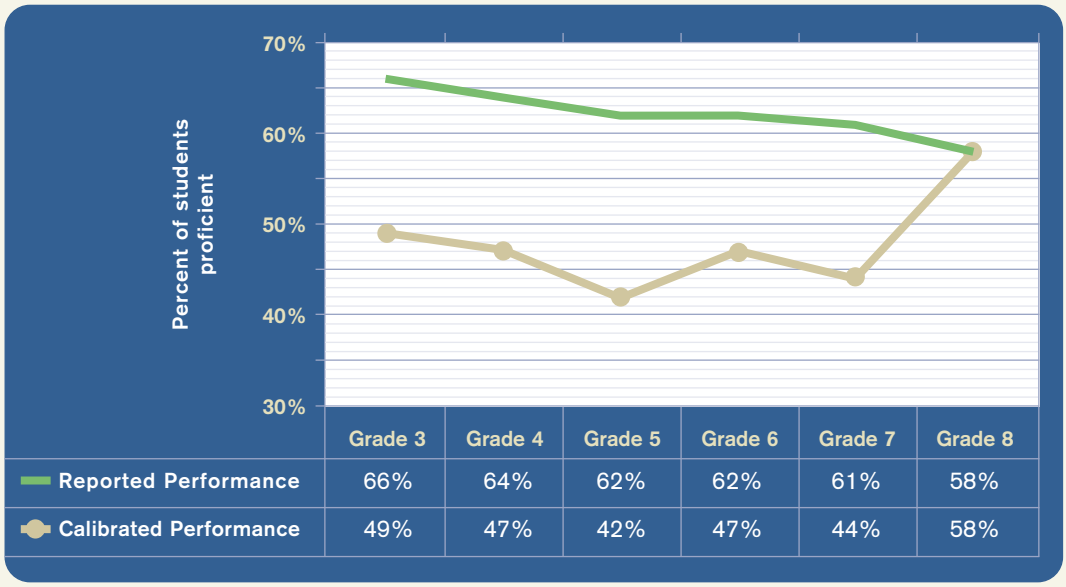
Figure 5 – Montana Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	81%	80%	79%	78%	77%	76%
Calibrated Performance	71%	69%	70%	72%	73%	76%

Note: This graphic shows, for example, that if Montana's grade-3 reading cut score were set at the same level of difficulty as its grade-8 cut score, 71 percent of third graders would achieve the proficient level, rather than 81 percent, as was reported by the state.

Figure 6 – Montana Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



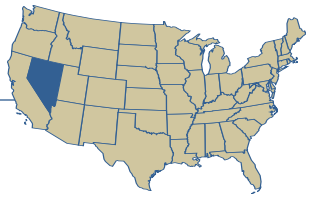
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	66%	64%	62%	62%	61%	58%
Calibrated Performance	49%	47%	42%	47%	44%	58%

Note: This graphic shows, for example, that if Montana's grade-3 mathematics cut score were set at the same level of difficulty as its grade-8 cut score, 49 percent of third graders would achieve the proficient level, rather than 66 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what it takes for a student to be considered proficient, Montana is relatively high for mathematics and in the middle of the pack for reading, compared with the other states in the study. In recent years, the state has adjusted the difficulty of these cut scores—making them more challenging in mathematics in eighth grade, and less challenging in both reading and math in fourth grade. As a result, Montana's expectations are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. Montana policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

Nevada



Introduction

This study linked data from the 2003 and 2006 administrations of Nevada's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Nevada's definitions of proficiency in reading and mathematics are relatively difficult at the early grades and about at the mid-point in the later grades, when compared to the 25 other states in the study. In other words, Nevada's tests are above average in terms of difficulty in the earlier grades and about average in the later grades.

The difficulty level of Nevada's tests remained constant from 2003 to 2006, except for a decline in third-grade reading expectations. Nonetheless, one striking finding of this study is that Nevada's cut scores are more difficult, relatively speaking, for third-grade students than they are for eighth-grade pupils. (In most states studied, the opposite is true.) Nevada policy-makers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Nevada Criterion-Referenced Assessment (Nevada CRT) and Iowa Test of Basic Skills (ITBS)

Nevada currently uses the Nevada Criterion-Referenced Assessment (Nevada CRT), which tests mathematics and reading in grades 3, 5, and 8, and the Iowa Test of Basic Skills (ITBS), which tests math, reading, language, and science in grades 4, 7, and 10. The same tests were used in spring 2003 in mathematics and reading: Nevada CRT in grades 3 and 5, and ITBS in grades 4 and 7. The current study linked reading and math data from spring 2003 and spring 2006 administrations to a common scale also administered in the 2003 and 2006 school years.

To determine the difficulty of Nevada's proficiency cut scores, we linked data from Nevada's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered "proficient.") This was done by analyzing a group of schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance was compared.)

Part 1: How Difficult are Nevada's Definitions of Proficiency in Reading and Math?

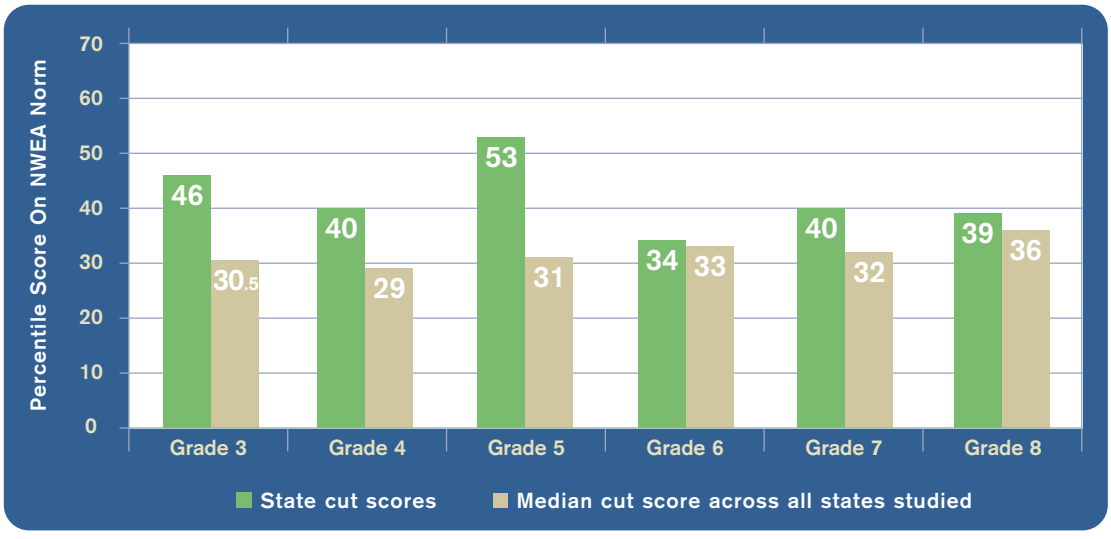
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach, we evaluated the difficulty of Nevada's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Nevada cut score on a test of equivalent difficulty. The two figures that follow show the difficulty of Nevada's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Nevada ranged between the 34th and 53rd percentiles for the norm group, with fifth grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 35th and 50th percentiles, with third grade being most challenging.

Nevada's reading cut scores are consistently above the median difficulty level, compared to the other states studied. For mathematics, Nevada's cut scores are above the median difficulty in grades 3 through 5 and below the median difficulty in grades 6 through 8.

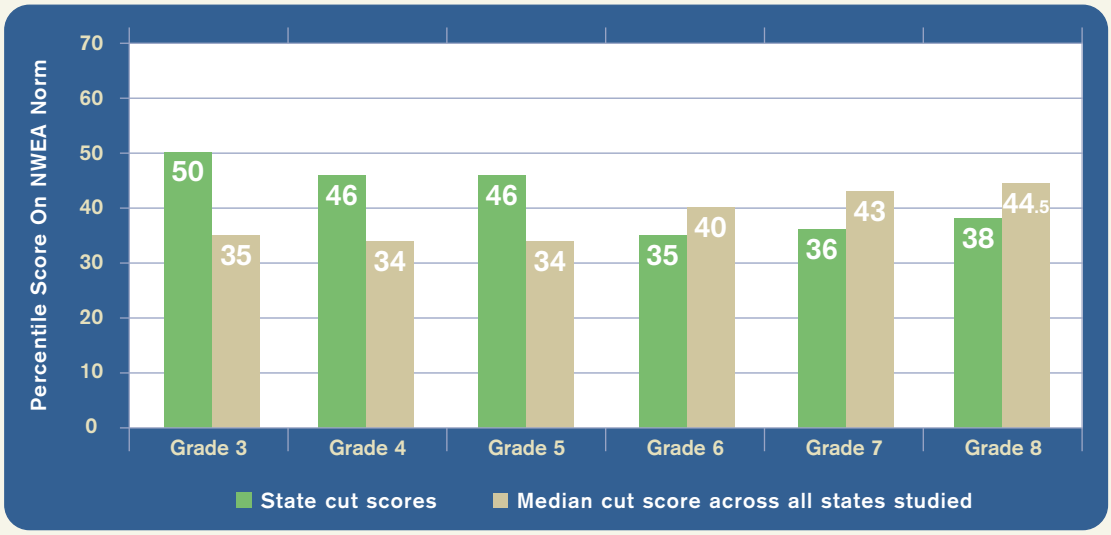
Another way of assessing difficulty is to evaluate how Nevada's proficiency cut scores rank relative to other states. Table 1 shows that the Nevada cut scores generally rank in the upper third in difficulty among the 26 states studied for this report. Its reading cut score in grades 3 and 5 and and math cut scores in grade 3 are particularly highly ranked: among the top two or three states in difficulty.

Figure 1 – Nevada Reading Cut Scores in Relation to All 26 States Studied, 2006 (as Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Nevada's cut scores are 1 to 22 percentile points above the median.

Figure 2 – Nevada Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (as Expressed in MAP Percentiles)



Note: Nevada's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of the 26 states reviewed in this study. The cut scores are 12 to 15 percentile points above the median in grades 3 through 7 and 5 to 7 percentile points below the median in grades 6 through 8.

Table 1 – Nevada Reading and Mathematics Proficiency Cut Scores Among 26 States for Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	3	5	2	12	7	8
Mathematics	3	5	8	16	18	14

Note: This table ranks Nevada's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Nevada's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2003 and 2006 school years. Cut score estimates for reading and mathematics were available for both years for grades 3 and 5.

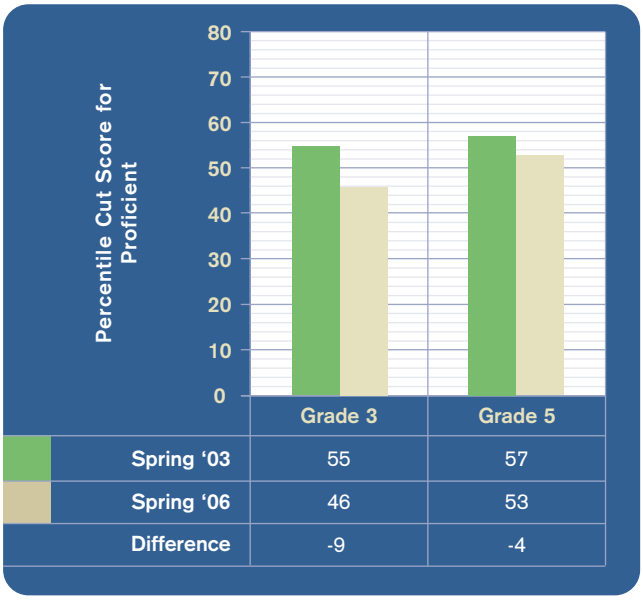
States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the exams used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Unintentional drift can occur even in states, such as Nevada, that maintained their proficiency levels.

Is it possible, then, to compare the proficiency scores between earlier administrations of Nevada tests and today's? Yes. Assume that we're judging a group of fifth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. Nevada CRT and ITBS in 2003 and Nevada CRT and ITBS in 2006 both can be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut scores needed to pass the Nevada CRT and ITBS in 2003 and 2006 on the MAP scale and ascertain whether the state's tests may have changed in difficulty.

Nevada's estimated **reading** cut scores showed a moderate decrease over this period in the third grade (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the third-grade reading proficiency rate in 2006 to be 9 percent higher than in 2003. (Nevada reported a 3-point gain for third graders over this period.) The proficiency cut score for fifth-grade reading remained essentially unchanged, as were all estimated **mathematics** cut scores (see Figure 4).

Thus, one could fairly say that Nevada's third-grade reading test was easier to pass in 2006 than in 2003, while the other tests stayed about the same. As a result, some apparent improvements in state-reported third-grade reading proficiency rate during this period may not be entirely a product of improved achievement.

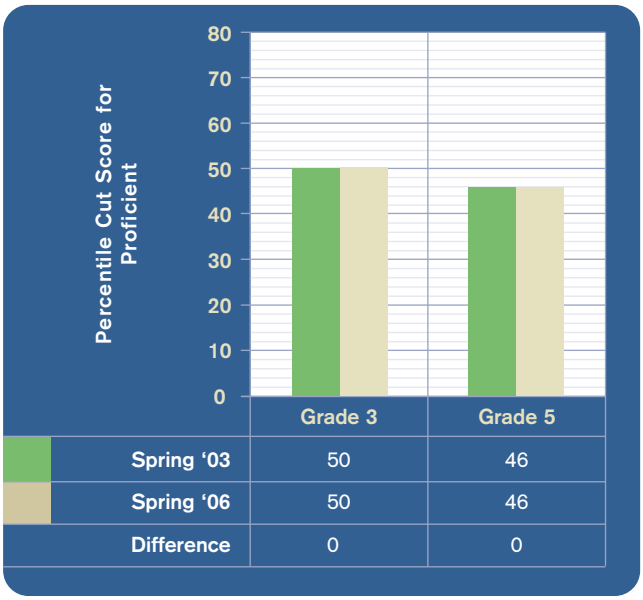
Figure 3 – Estimated Difference in Nevada's Proficiency Cut Scores in Reading, 2003-2006 (as Expressed in MAP Percentiles)



	Grade 3	Grade 5
Spring '03	55	57
Spring '06	46	53
Difference	-9	-4

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third-grade students in 2003 had to score at the 55th percentile on NWEA norms in order to be considered proficient, while in 2006 third graders had only to score at the 46th percentile to achieve proficiency. The changes in grade 5 were within the margin of error (in other words, too small to be considered substantive).

Figure 4 – Estimated Difference in Nevada's Proficiency Cut Scores in Mathematics, 2003-2006 (as Expressed in MAP Percentiles)



	Grade 3	Grade 5
Spring '03	50	46
Spring '06	50	46
Difference	0	0

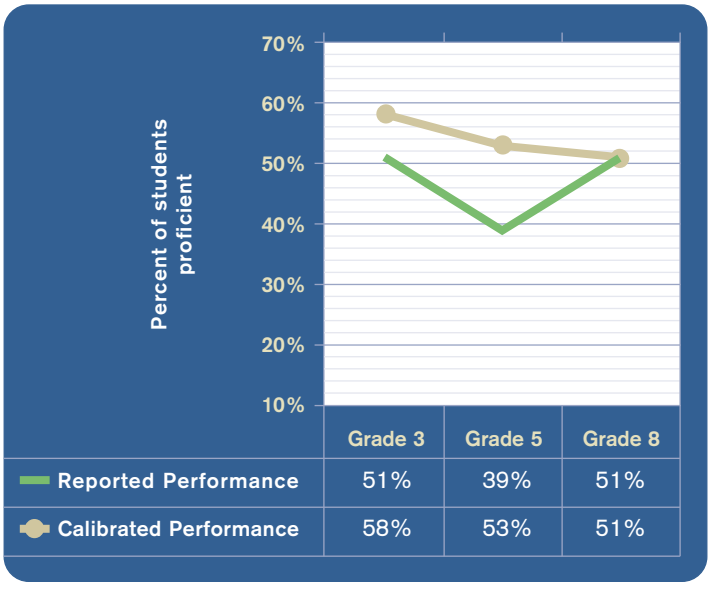
Note: This graphic shows that the difficulty of achieving proficiency in math has not changed. For example, third-grade students in both 2003 and 2006 had to score at the 50th percentile on NWEA norms in order to be considered proficient. The changes in grades 3 and 5 were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth-grade cut score is no more or less difficult for eighth graders to achieve than the third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

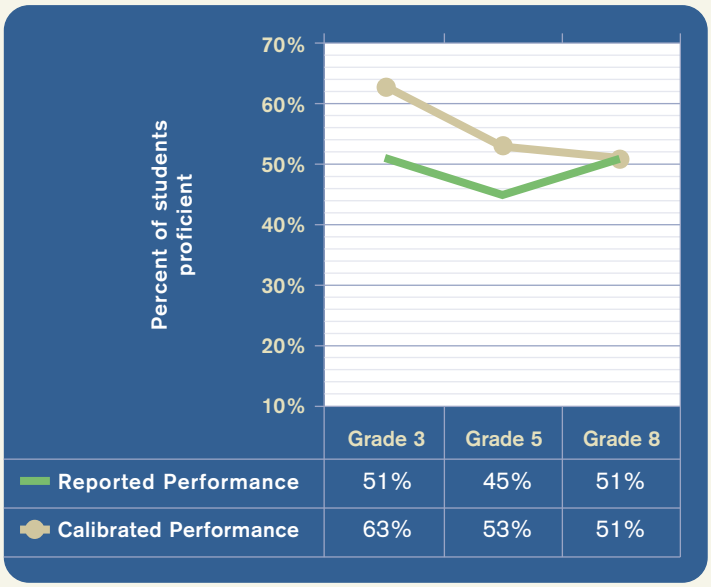
Figures 1 and 2 illustrated the relative difficulties of Nevada's cut scores for reading and mathematics, showing that the upper-grade cut scores in reading and mathematics were less challenging than in the lower grades. The following two figures show Nevada's reported performance in reading (Figure 5) and mathematics (Figure 6) on the state test and the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent across grades. This would lead to the conclusion that the more difficult standards at the lower grades may result in underestimating the proportion of third-grade students who are actually on track to meet the easier proficiency standards of the later grades.

Figure 5 – Nevada Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



Note: This graphic shows, for example, that, if Nevada's grade-3 reading cut score were set at the same level of difficulty as its grade-8 cut score, 58 percent of third graders would achieve the proficient level, rather than 51 percent, as was reported by the state.

Figure 6 – Nevada Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



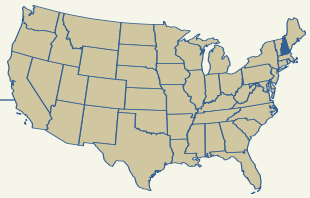
	Grade 3	Grade 5	Grade 8
Reported Performance	51%	45%	51%
Calibrated Performance	63%	53%	51%

Note: This graphic shows, for example, that, if Nevada's grade-3 mathematics cut score were set at the same level of difficulty as its grade-8 cut score, 63 percent of third graders would achieve the proficient level, rather than 51 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what students should know and be able to do in order to be considered proficient in reading and math, Nevada is relatively high at the lower grades and at about the mid-point for the upper grades, at least compared to the other 25 states in this study. This finding is roughly consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which found Nevada's standards to be in the upper half for the early grades. In recent years, the difficulty of the third-grade reading cut score has decreased while other tests and grades have held roughly constant. Furthermore, Nevada's proficiency cut scores are not smoothly calibrated across grades; some students who are not proficient in third grade actually may be on track to be proficient by the eighth grade. Nevada policymakers might consider adjusting their cut scores across grades so that performance at the early grades accurately predicts proficiency at the higher grades.

New Hampshire



Introduction

This study linked data from the 2003 and 2005 administrations of New Hampshire's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that New Hampshire's definitions of proficiency in reading and mathematics are relatively consistent with the standards set by the other 25 states in this study, with its reading and math tests a bit above average in difficulty.

The difficulty of New Hampshire's tests increased markedly from 2003 to 2005—the No Child Left Behind era—from very low to moderate standards. The state's cut scores are also now less challenging for third-grade students than for eighth graders. New Hampshire policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: New Hampshire - New England Common Assessment Program (NECAP)

New Hampshire currently uses an assessment called the New England Common Assessment Program (NECAP) which tests mathematics and reading in grades 3-8. It replaced the New Hampshire Educational Improvement and Assessment Program (NHEIAP) that was used prior to fall 2005 and that tested math and reading in students in grades 3, 6, and 10. The current study linked data from fall 2003 and fall 2005 administrations to a common scale that was also administered in the 2003 and 2005 school years.

To determine the difficulty of New Hampshire's proficiency cut scores, we linked reading and math data from New Hampshire's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are New Hampshire's Definitions of Proficiency in Reading and Math?

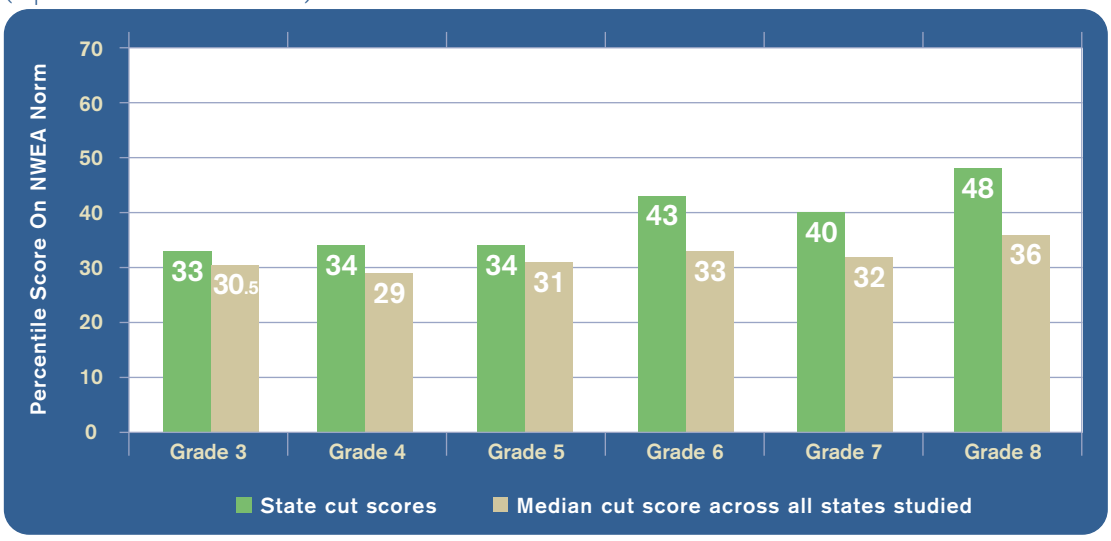
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high jump bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of New Hampshire's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the New Hampshire cut score on a test of equivalent difficulty. The following two figures show the difficulty of New Hampshire's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in New Hampshire ranged between the 33rd and 48th percentiles for the norm group, with the eighth grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 34th and 53rd percentiles, with eighth grade again being most challenging.

New Hampshire's cut scores in both reading and mathematics are consistently at or above the median in difficulty among the states studied. Note, though, that New Hampshire's cut scores for reading are generally lower than for math at the same grade. (This was the case in the majority of states studied.) Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, New Hampshire students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentages that pass state tests in those subjects.

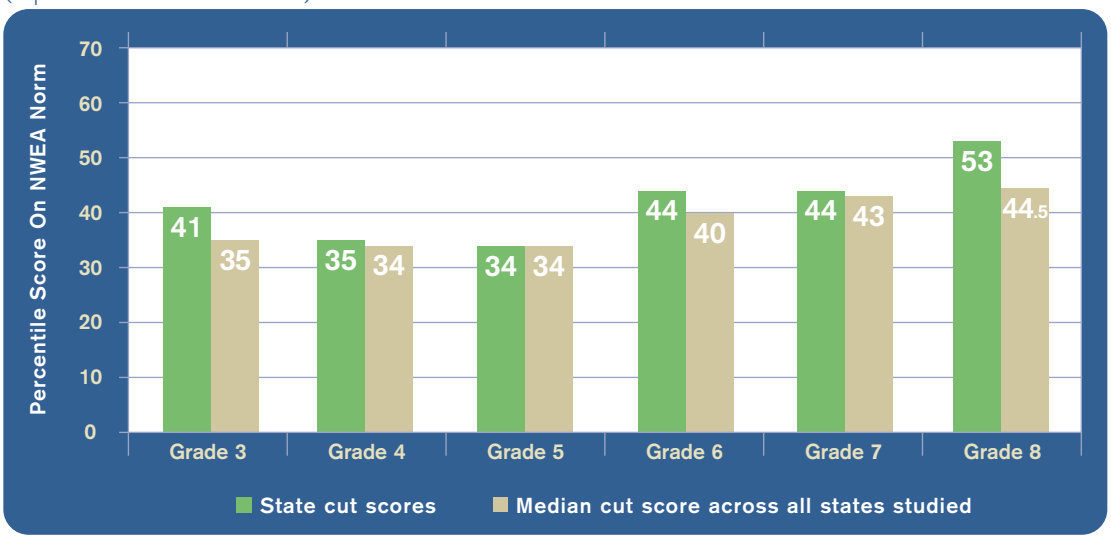
Another way of assessing difficulty is to evaluate how New Hampshire's proficiency cut scores rank relative to other states. Table 1 shows that the New Hampshire cut scores generally rank in the upper third for reading and around the middle for math, among the 26 states studied for this report. Its reading cut score in grade eight is particularly high, ranking third out of the 26 states.

Figure 1 – New Hampshire Reading Cut Scores in Relation to All 26 States Studied, 2005 (expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. New Hampshire's cut scores are consistently 2.5 to 12 percentile points above the median.

Figure 2 – New Hampshire Mathematics Cut Scores in Relation to All 26 States Studied, 2005 (expressed in MAP Percentiles)



Note: New Hampshire's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. The state's cut scores are consistently 1 to 8.5 percentile points above the median, with the exception of grade 5 where it matches the median.

Table 1 – New Hampshire Rank Among 26 States for Proficiency Cut Scores in Reading and Mathematics, 2005

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	9	6	7	4	7	3
Mathematics	8	10	13	9	9	6

Note: This table ranks New Hampshire's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Differences in Cut Scores over Time

In order to measure their consistency, New Hampshire's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2003-4 and 2005-6 school years. Cut score estimates for reading and math were available for both years in grades 3 and 6.

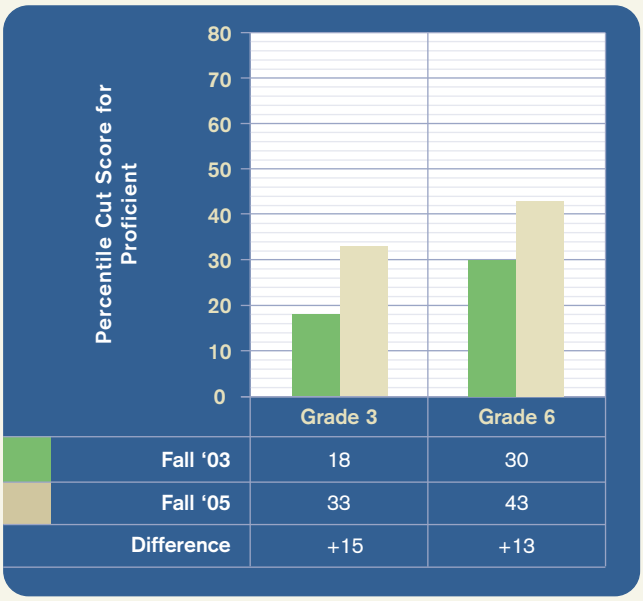
States may periodically re-adjust the cut scores they use to define proficiency in reading and mathematics, or, as New Hampshire did, may change or update the tests used to test student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed.

Is it possible, then, to compare the proficiency scores between earlier administrations of New Hampshire tests and today's? Yes. Assume that we're judging a group of fifth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. Although the NHEIAP and NECAP are different measures, both can be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the NHEIAP in 2003 and the NECAP in 2005 and ascertain which test was more difficult. It should be noted, however, that for the NHEIAP in 2003, the "basic" level was the minimum satisfactory performance level reported by New Hampshire for purposes of NCLB, whereas when the NECAP was adopted, the "proficient" level became the minimum acceptable level reported for NCLB. Furthermore, the NHEIAP administered in 2003 was a spring season test, and the NECAP is a fall test. These changes in practice are accounted for in the following analyses and figures.

New Hampshire's estimated **reading** cut scores indicate large increases over this two-year period in the third and sixth grades (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the reading proficiency rates in 2005 to be 15 and 13 points lower than in 2003 for third and sixth graders, respectively. (New Hampshire reported a 4 point drop for third graders and a 9 point drop for sixth graders over this period.)

New Hampshire's estimated **mathematics** cut scores show similar patterns, with large increases for grades 3 and 6 (Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the math proficiency rate in 2005 to be 35 percent lower than in 2003 for third grade, and 22 percent lower for sixth grade. (New Hampshire reported a 16-point drop for third graders and a 12-point drop for sixth graders over this period.) Thus, one could fairly say that New Hampshire's reading and mathematics tests were harder to pass in 2005 than in 2003, at least at the third and sixth grades.

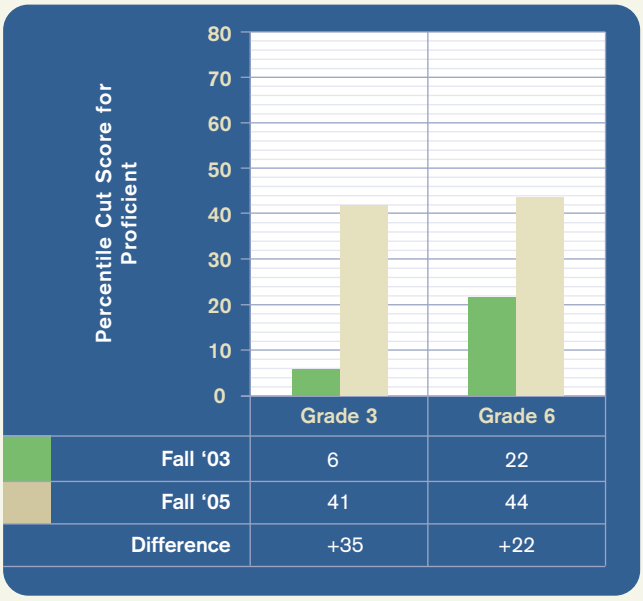
Figure 3 – Estimated Differences in New Hampshire's Proficiency Cut Scores in Reading, 2003-2005 (as Expressed in MAP Percentiles)



	Grade 3	Grade 6
Fall '03	18	30
Fall '05	33	43
Difference	+15	+13

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, New Hampshire sixth grade students in 2003 had to score at the 30th percentile on NWEA norms in order to be considered proficient, while by 2005 sixth graders had to score at the 43rd percentile to achieve proficiency.

Figure 4 – Estimated Differences in New Hampshire's Proficiency Cut Scores in Mathematics, 2003-2005 (as Expressed in MAP Percentiles)



	Grade 3	Grade 6
Fall '03	6	22
Fall '05	41	44
Difference	+35	+22

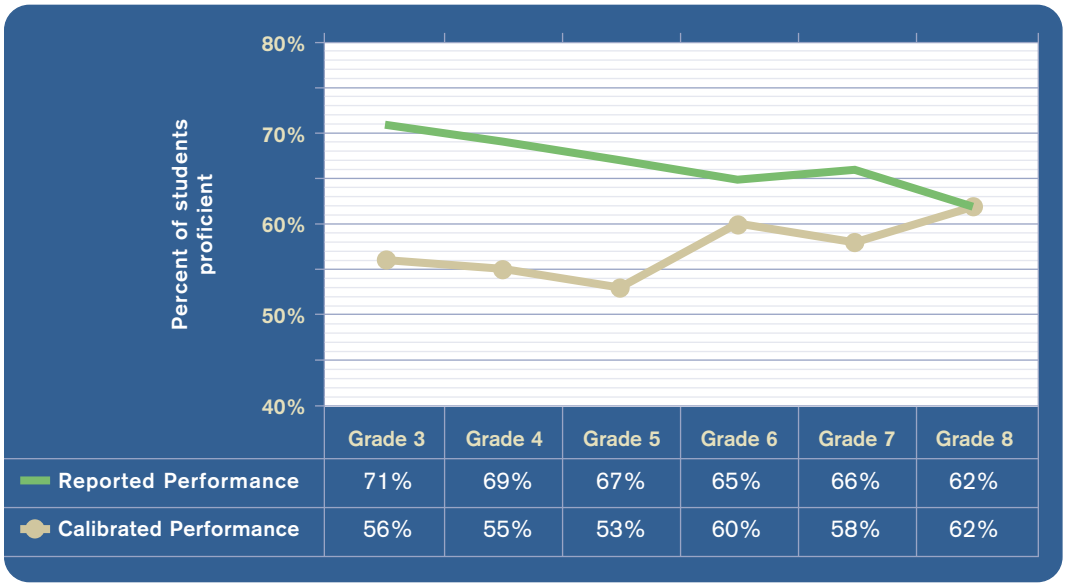
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, third grade students in 2003 had to score at the 6th percentile nationally in order to be considered proficient, while in 2005 sixth graders had to score at the 41st percentile to achieve proficiency.

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining New Hampshire's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed the relative difficulty of New Hampshire's reading and mathematics cut scores across the different grades, indicating that that the upper grade cut scores in both subjects were somewhat more challenging than in the lower grades. (This was the case for the majority of states studied.) The following two figures show New Hampshire's reported 2005 performance in reading (Figure 5) and mathematics (Figure 6) on its state test and the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-eight standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades. This would lead to the conclusion that the higher rates of proficiency that the state has reported for lower grades students are somewhat misleading.

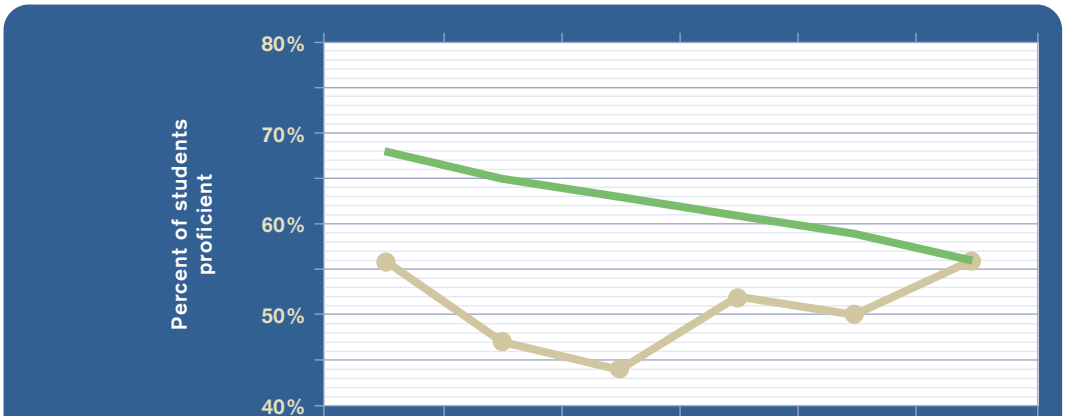
Figure 5 – New Hampshire Reading Performance as Reported and as Calibrated to the Grade-8 Standard, fall 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	71%	69%	67%	65%	66%	62%
Calibrated Performance	56%	55%	53%	60%	58%	62%

Note: This graphic shows, for example, that if New Hampshire's grade-3 reading cut score was set at the same level of difficulty as its grade-8 cut score, 56 percent of third graders would achieve the proficient level, rather than 71 percent, as was reported by the state.

Figure 6 – New Hampshire Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, fall 2005



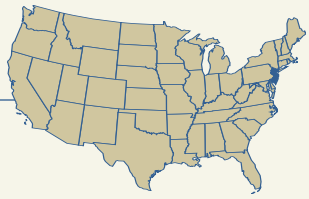
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	68%	65%	63%	61%	59%	56%
Calibrated Performance	56%	47%	44%	52%	50%	56%

Note: This graphic shows, for example, that if New Hampshire's grade-3 mathematics cut score were set at the same level of difficulty as its grade-8 cut score, 56 percent of third graders would achieve the proficient level, rather than 68 percent, as was reported by the state.

Policy Implications

When determining what constitutes proficiency in reading and math, New Hampshire is just above the middle of the pack, at least compared with the other 25 states in this study. However, New Hampshire increased its cut scores dramatically from their previous levels when it adopted the New England Common Assessment Program. Also of note is that New Hampshire's cut scores are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by eighth grade. State policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

New Jersey



Introduction

This study linked data from the 2005 and 2006 administrations of New Jersey's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that New Jersey's definitions of proficiency in reading and mathematics are less difficult than the cut scores set by the majority of the other 25 states in this study, at least in the lower grades. In other words, New Jersey's tests are generally below average in terms of difficulty.

The level of difficulty changed some from 2005 to 2006, but the direction of that change varied by grade and subject. New Jersey's reading tests have grown harder to pass, while the mathematics tests are now easier to pass, although not for all grades. One finding of this study is that New Jersey's cut scores are easier for third-grade students than for middle-school students (taking into account the differences in subject content and children's development). State policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: New Jersey Assessment of Knowledge and Skills (NJ ASK) and Grade Eight Proficiency Assessment (GEPA)

New Jersey currently uses an assessment called the New Jersey Assessment of Knowledge and Skills (NJ ASK), which tests language arts literacy and mathematics in students in grades three through seven, the New Jersey Grade Eight Proficiency Assessment (GEPA), which tests language arts literacy, mathematics, and science in students in grade eight, and the New Jersey High School Proficiency Assessment (HSPA), which tests language arts literacy and mathematics in students in grade 10. The same tests were used in spring 2005. The current study linked data from spring 2005 and spring 2006 NJ ASK and GEPA administrations to a common scale also administered in the 2005 and 2006 school years.

To determine the difficulty of New Jersey's proficiency cut scores, we linked data from New Jersey's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are New Jersey's Definitions of Proficiency in Reading and Math?

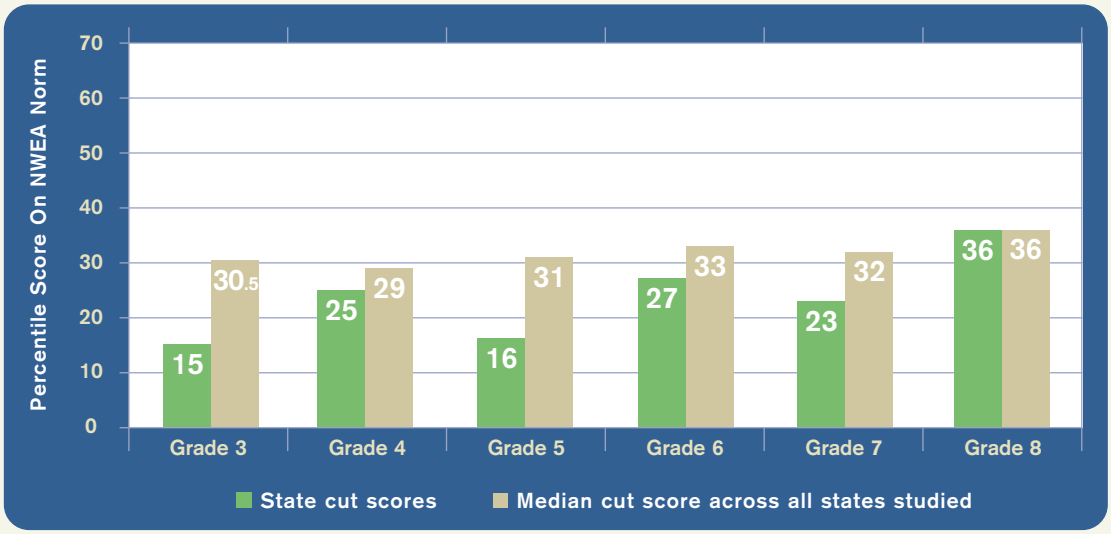
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of New Jersey's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the New Jersey cut score on a test of equivalent difficulty. The following two figures show the difficulty of New Jersey's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all states in the study. The proficiency cut scores for **reading** in New Jersey ranged between the 15th and 36th percentiles of the NWEA norm group, with eighth grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 13th and 43rd percentiles, with seventh grade being most challenging.

For most grades, New Jersey's reading cut scores fall below the median difficulty among the states studied. This is also true at the lower grades for mathematics, although the math cut scores in grades six and seven equal the median difficulty. Note, too, that in grades five, six, and seven, New Jersey's cut scores for reading are lower than those for mathematics. (This was the case in most grades in most states.) Thus, reported differences in achievement on the NJ ASK between reading and mathematics might be more a product of differences in cut scores than in actual student achievement. In other words, New Jersey students may be performing worse in reading, or better in math, in grades five through seven than is apparent by just looking at the percentage of students passing state tests in those subjects.

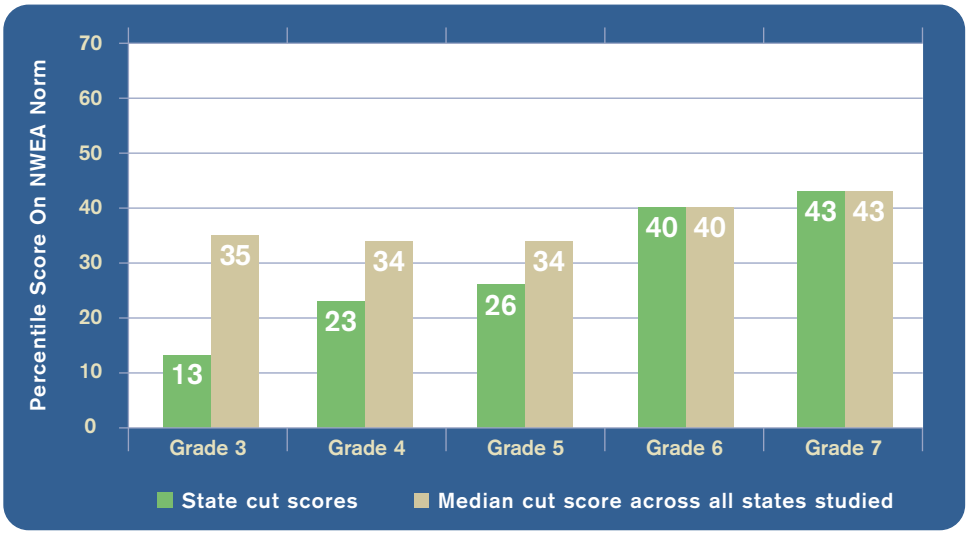
Another way of assessing difficulty is to evaluate how New Jersey's proficiency cut scores rank relative to other states. Table 1 shows that the New Jersey cut scores generally rank in the lower half in difficulty among the 26 states studied for this report, except for math in the upper grades and reading in grade eight. The standards set for grade-three reading and mathematics are among the lowest: 22nd and 20th of 26, respectively.

Figure 1 – New Jersey Reading Cut Scores in Relation to All 26 States Studied, 2006
(Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Only in eighth grade does New Jersey's cut score reach the median. Cut scores in grades three through seven are 4 to 15.5 percentile points below the median.

Figure 2 – New Jersey Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: New Jersey's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. Grades six and seven cut scores reach the median, but those in grades three through five fall 8 to 22 percentile points below the median.

Table 1 – New Jersey Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	22	17	23	18	22	9
Mathematics	23	22	18	12	12	Not Available

Note: This table ranks New Jersey's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Differences in Cut Scores over Time

In order to measure their consistency, New Jersey's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2005 and 2006 school years. Cut score estimates at both years were available in reading and mathematics for grades three and four.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Plus, unintentional drift can occur even in states, such as New Jersey, that maintained their proficiency levels.

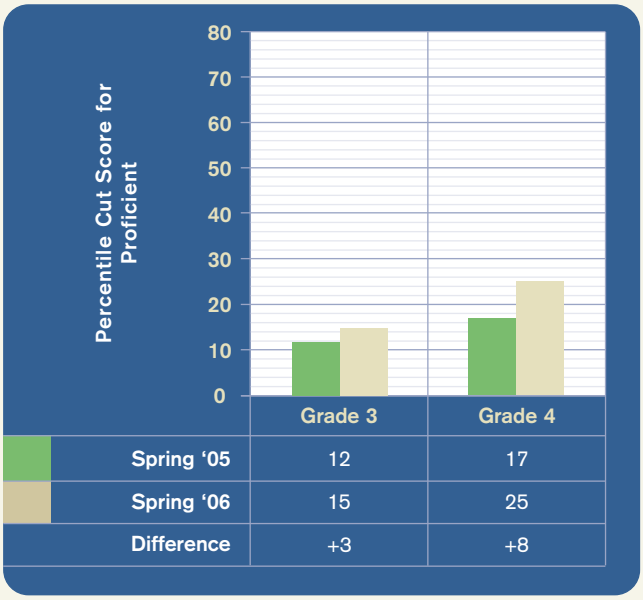
Is it possible, then, to make comparisons of the proficiency scores between earlier administrations of New Jersey tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measures or scales used by the NJ ASK in 2005 and in 2006 can both be linked to the scale that was used to report MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the NJ ASK in 2005 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty. This allows us to estimate whether the 2006 NJ ASK was easier to pass, more difficult, or about the same as in 2005.

New Jersey's estimated **reading** cut scores indicate increases over this duration in the third and fourth grade (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the reading proficiency rate in 2006 to be three percent lower in 2006 than in 2005 for third grade, and about eight percent lower in fourth grade. (New Jersey reported a 1-point drop for third graders and a 2-point drop for fourth graders over this period.)

New Jersey's estimated **mathematics** cut scores show a decrease in the difficulty at third grade (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, this would likely yield an increased proficiency rate of nine percent (see Figure 4). (New Jersey reported a 4-point gain for third graders over this period.) The fourth-grade mathematics proficiency cut score did not change substantively from its 2005 level.

Thus, one could fairly say that New Jersey's reading tests were harder to pass in 2006 than in 2005, while the mathematics test became easier to pass for third graders. As a result, improvements in the state's third-grade mathematics proficiency rate may not be entirely a product of improved achievement, while any actual improvements in reading performance may be masked somewhat by the increased difficulty of the state's proficiency cut scores.

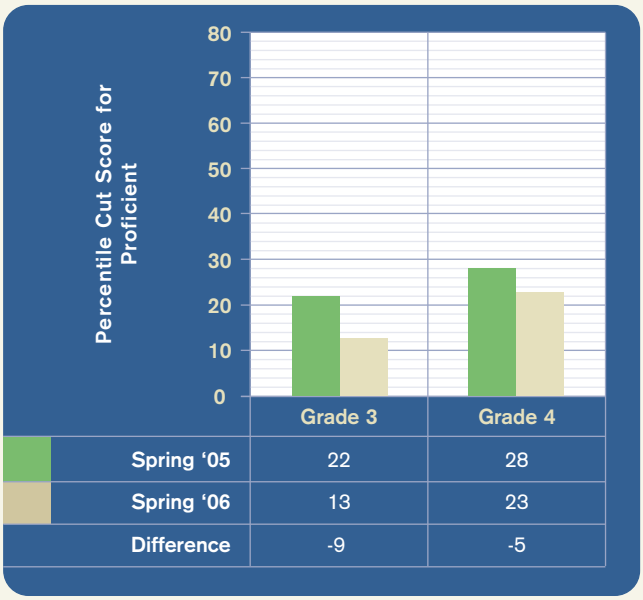
Figure 3 – Estimated Differences in New Jersey's Proficiency Cut Scores in Reading, 2005-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4
Spring '05	12	17
Spring '06	15	25
Difference	+3	+8

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third grade students in 2005 had to score at the 12th percentile on the NWEA norm in order to be considered proficient, while in 2006 third graders had to score at the 15th percentile to achieve proficiency.

Figure 4 – Estimated Differences in New Jersey's Proficiency Cut Scores in Mathematics, 2005-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4
Spring '05	22	28
Spring '06	13	23
Difference	-9	-5

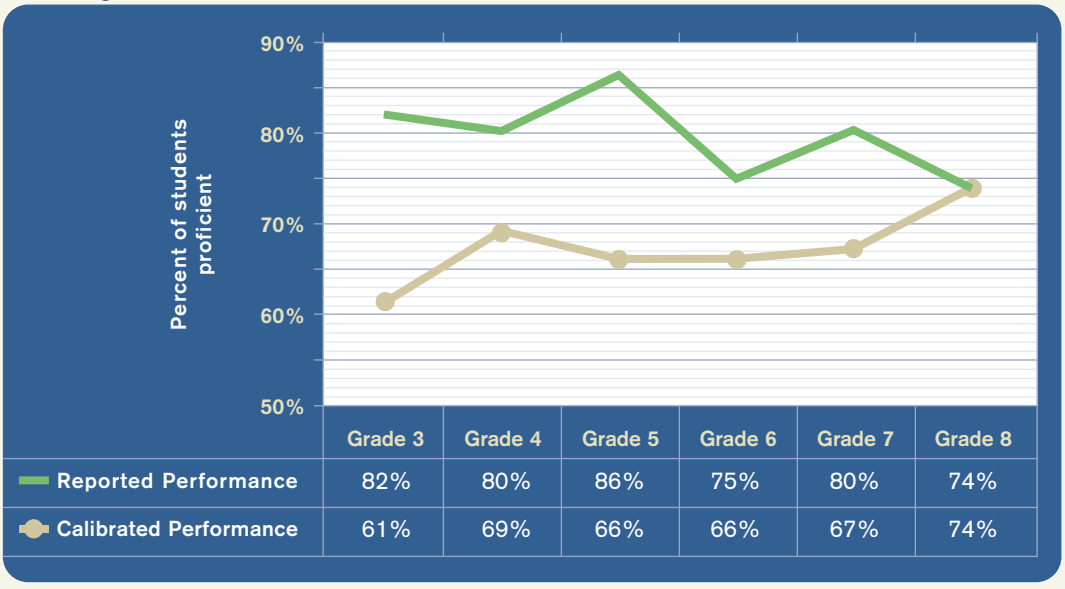
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, third-grade students in 2005 had to score at the 22nd percentile on the NWEA norm in order to be considered proficient, while a year later third graders had only to score at the 13th percentile to achieve proficiency. The changes in grade four were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth-grade cut score is no more or less difficult to achieve for eighth graders than the third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the cut scores at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 illustrated the relative difficulty of the reading and mathematics cut scores across grades, showing that the upper-grade cut scores in reading and mathematics were more difficult than the cut scores in the lower grades. The two figures that follow show New Jersey's reported performance in reading (Figure 5) and mathematics (Figure 6) on the state test, compared with the rates of proficiency that would be achieved if the cut scores were all calibrated to the grade-seven standard (in math) or grade-eight standard (in reading). When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades. This would lead to the conclusion that the higher rates of proficiency that the state has reported for students in the earlier grades are somewhat misleading.

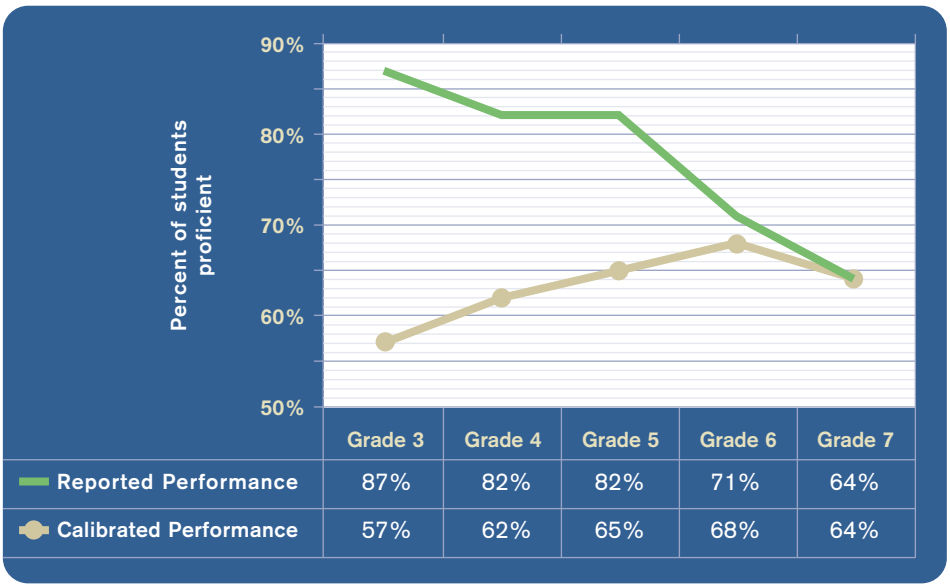
Figure 5 – New Jersey Reading Performance as Reported and as Calibrated to the Grade-Eight Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	82%	80%	86%	75%	80%	74%
Calibrated Performance	61%	69%	66%	66%	67%	74%

Note: This graphic shows, for example, that if New Jersey's grade-three reading cut score was set at the same level of difficulty as its grade-eight cut score, 61 percent of third graders would achieve the proficient level, rather than 82 percent, as was reported by the state.

Figure 6 – New Jersey Mathematics Performance as Reported and as Calibrated to the Grade-Seven Standard, 2006



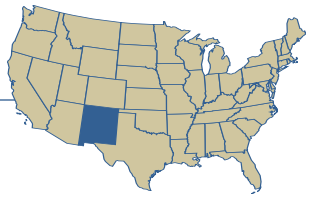
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Reported Performance	87%	82%	82%	71%	64%
Calibrated Performance	57%	62%	65%	68%	64%

Note: This graphic shows, for example, that if New Jersey's grade-three mathematics cut score was set at the same level of difficulty as its grade-seven cut score, 57 percent of third graders would achieve the proficient level, rather than 87 percent, as was reported by the state.

Policy Implications

When setting cut scores for what it takes for a student to be considered proficient in reading and math, New Jersey is relatively low, particularly in the earlier grades, at least compared to the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found New Jersey's standards to be in the bottom half of the state distribution for the earlier grades (though slightly higher for the upper grades). From 2005 to 2006, New Jersey's proficiency cut scores changed somewhat, becoming more challenging for reading and somewhat easier for mathematics – though not for all grades. Plus, New Jersey's cut scores are not calibrated smoothly across grades; students who are proficient in third grade are not necessarily on track to be proficient by the end of middle school. New Jersey policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

New Mexico



Introduction

This study linked data from the 2005 and 2006 administrations of New Mexico's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that New Mexico's definitions of proficiency in reading are consistent with the cut scores set by the 25 other states in this study, while its definitions for mathematics proficiency are relatively more difficult. In other words, New Mexico's reading tests are about average in terms of difficulty, while its math tests are above average.

However, the level of difficulty of New Mexico's math tests declined somewhat from 2005 to 2006, although not for all grades. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the New Mexico test not being matched by learning gains on the Northwest Evaluation Association test. Additionally, New Mexico's mathematics cut scores are now relatively less difficult for third-grade students than they are for eighth-grade students (taking into account the differences in subject content and children's development). State policymakers might consider adjusting their math cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

What We Studied: New Mexico Standards Based Assessments (NMSBA)

New Mexico currently uses an assessment called the New Mexico Standards Based Assessments (NMSBA) which tests mathematics, language arts, and science in students in grades three through nine and math and language arts in grade 11. The tests were used in spring 2005. The current study linked reading and math data from spring 2005 and spring 2006 test administrations to a common scale also administered in the 2005 and 2006 school years.

To determine the difficulty of New Mexico's proficiency cut scores, we linked data from NMSBA to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are New Mexico's Definitions of Proficiency in Reading and Math?

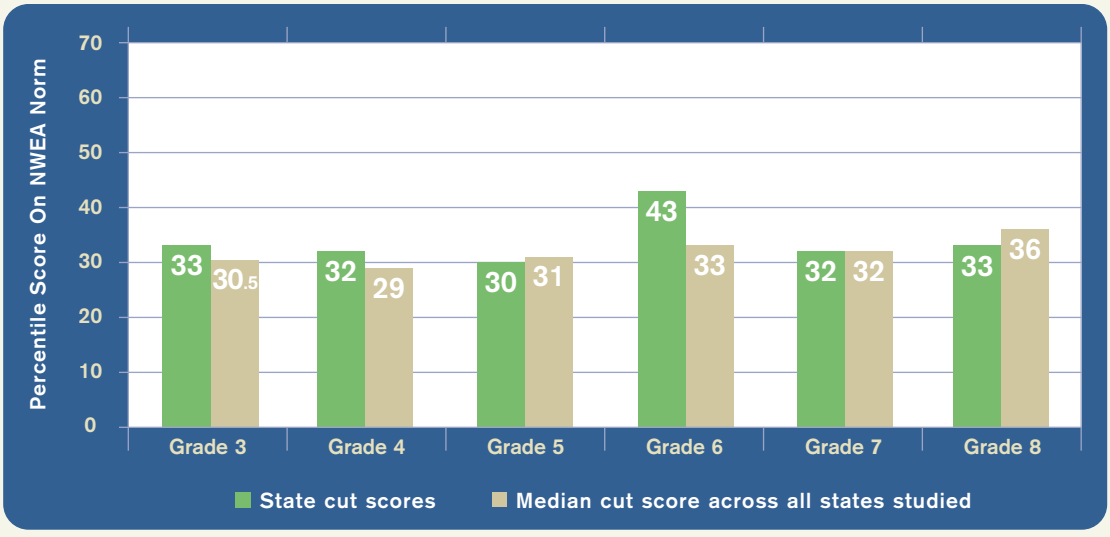
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of New Mexico's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the New Mexico cut score on a test of equivalent difficulty. The following two figures show the difficulty of New Mexico's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in New Mexico ranged between the 30th and 43rd percentiles nationally, with sixth grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 46th and 61st percentiles, with seventh grade being most challenging.

Except in grade six, New Mexico's reading cut scores are near the median difficulty of the states studied, whereas New Mexico's mathematics cut scores are higher than the median in all grades. Note, too, that New Mexico's cut scores for reading are lower than the cut scores for mathematics. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, New Mexico students may be performing worse in reading and/or better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

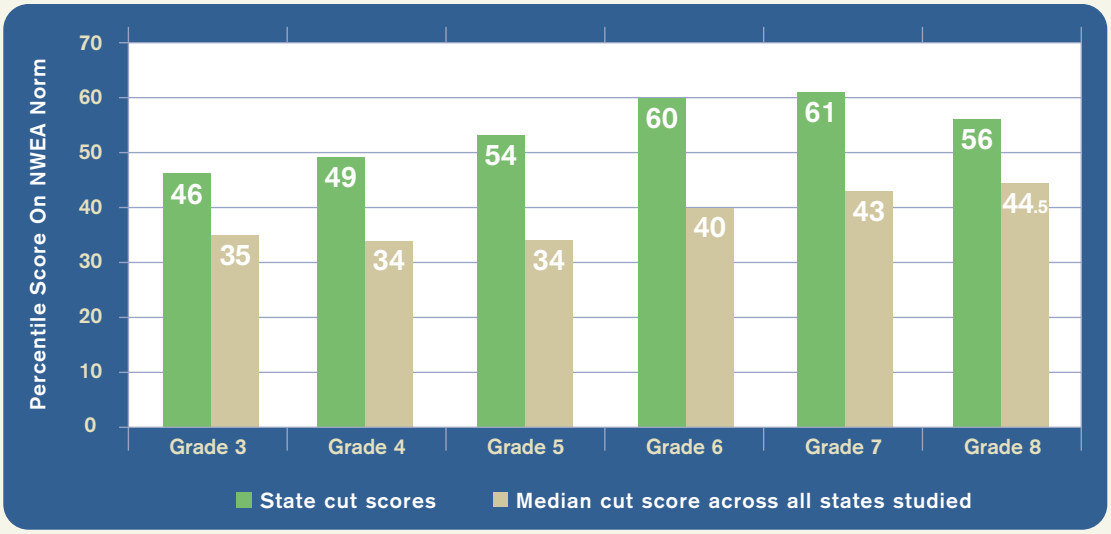
Another way of assessing difficulty is to evaluate how New Mexico's proficiency cut scores rank relative to other states. Table 1 shows that the New Mexico reading cut scores generally rank in the top half in difficulty while math cut scores rank among the top three or four states in every grade.

Figure 1 – Estimates of New Mexico Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. New Mexico's reading cut scores hover around the median, with the exception of grade 6, in which the state cut score is 10 percentile points higher.

Figure 2 – Estimates of New Mexico Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: New Mexico's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. Across grades, New Mexico's math cut scores are above the median.

Table 1 – New Mexico Rank for Proficiency Cut Scores in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	9	10	14	4	13	14
Mathematics	4	4	4	4	3	4

Note: This table ranks New Mexico's cut scores relative to the cut scores of the other 25 states in the study, with 1 being the highest and 26 the lowest.

Part 2: Differences in Cut Scores over Time

In order to measure their consistency, New Mexico's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2005 and 2006 school years. Cut score estimates for reading and mathematics were available for both years in grades three through eight.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Plus, unintentional drift can occur even in states, such as New Mexico, that maintained their proficiency levels.

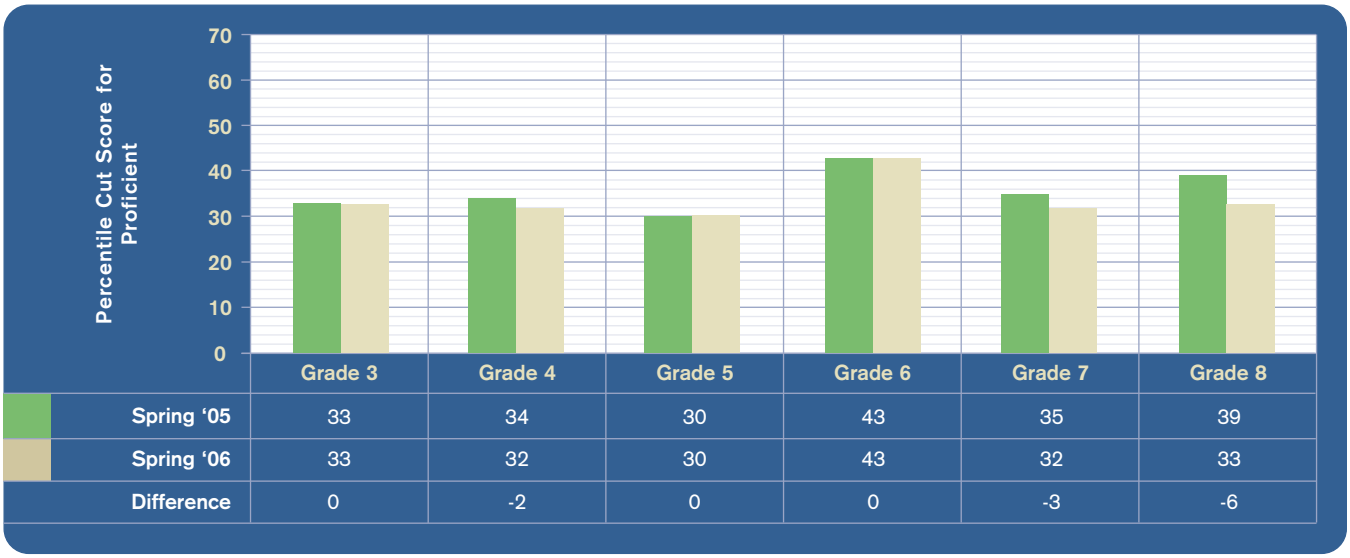
Is it possible, then, to make comparisons of the proficiency scores between earlier administrations of New Mexico tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The 2005 and 2006 NMSBA can both be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut scores needed to pass the NMSBA in 2005 and 2006 on the MAP scale and ascertain whether the state test may have changed in difficulty.

New Mexico's estimated **reading** cut scores indicate no substantive changes over this one-year period (see Figure 3). Consequently, one would expect that any changes in the reported reading proficiency ratings could be directly attributable to actual changes in student performance.

New Mexico's estimated **mathematics** cut scores show substantive decreases for grades six and eight (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, this would likely yield increases of seven and six percent, respectively, in the state-reported mathematics proficiency rates for those grades. (New Mexico reported a 2-point gain for sixth graders and a 2-point gain for eighth graders over this period.)

Thus, one could fairly say that New Mexico's reading tests remained about the same from 2005 to 2006, but that math tests for grades six and eight became easier to pass. As a result, some apparent improvements in the state's sixth- and eighth-grade mathematics proficiency rates during this period may not be entirely a product of improved achievement.

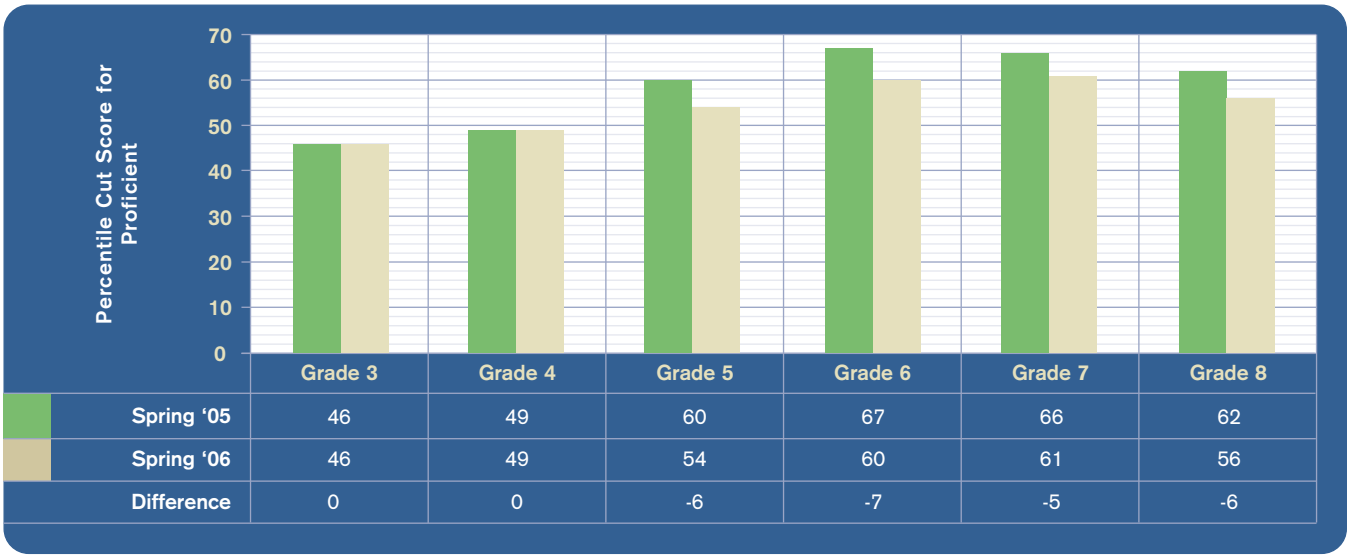
Figure 3 – Estimated Change in New Mexico's Proficiency Cut Scores in Reading, 2005-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Spring '05	33	34	30	43	35	39
Spring '06	33	32	30	43	32	33
Difference	0	-2	0	0	-3	-6

Note: This graphic shows that the difficulty of achieving proficiency in reading has not changed. For example, third-grade students in 2005 had to score at the 33rd percentile on NWEA norms in order to be considered proficient, and in 2006 third graders still had to score at the 33rd percentile to achieve proficiency. The observed changes in all grades were within the margin of error (in other words, too small to be considered substantive).

Figure 4 – Estimated Difference in New Mexico's Proficiency Cut Scores in Mathematics, 2005-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Spring '05	46	49	60	67	66	62
Spring '06	46	49	54	60	61	56
Difference	0	0	-6	-7	-5	-6

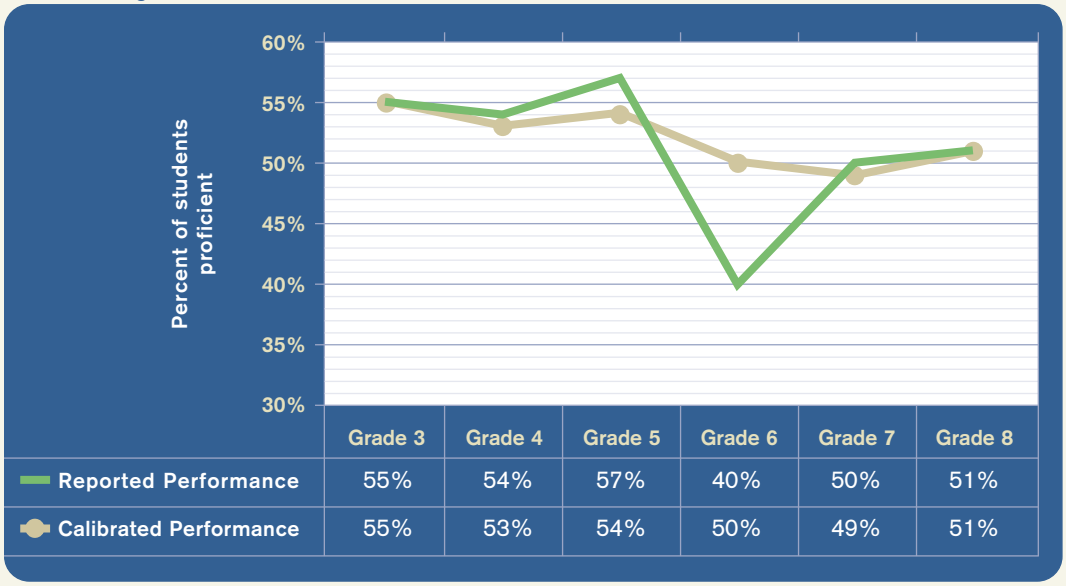
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, sixth-grade students in 2005 had to score at the 67th percentile on NWEA norms in order to be considered proficient, while in 2006 sixth graders had only to score at the 60h percentile to achieve proficiency. The changes in grades three, four, five, and seven were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth-grade cut score is no more or less difficult to achieve for eighth graders than the third-grade cut scores is for third graders, respectively. When cut scores are all calibrated, to the grade-eight standard, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the cut scores at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 indicated the relative difficulty of the reading and math cut scores, showing that reading cut scores were consistent except in grade four, which was relatively more difficult. In mathematics, however, cut scores were less difficult in the lower grades than in the upper. (This pattern held true for most states studied.) The two figures that follow show New Mexico's reported performance in reading (Figure 5) and mathematics (Figure 6) on the state test, compared with the rates of proficiency that would be achieved if the cut scores were calibrated across grades. When grade-to-grade differences in difficulty of the cut score are removed, student performance is more consistent at all grades.

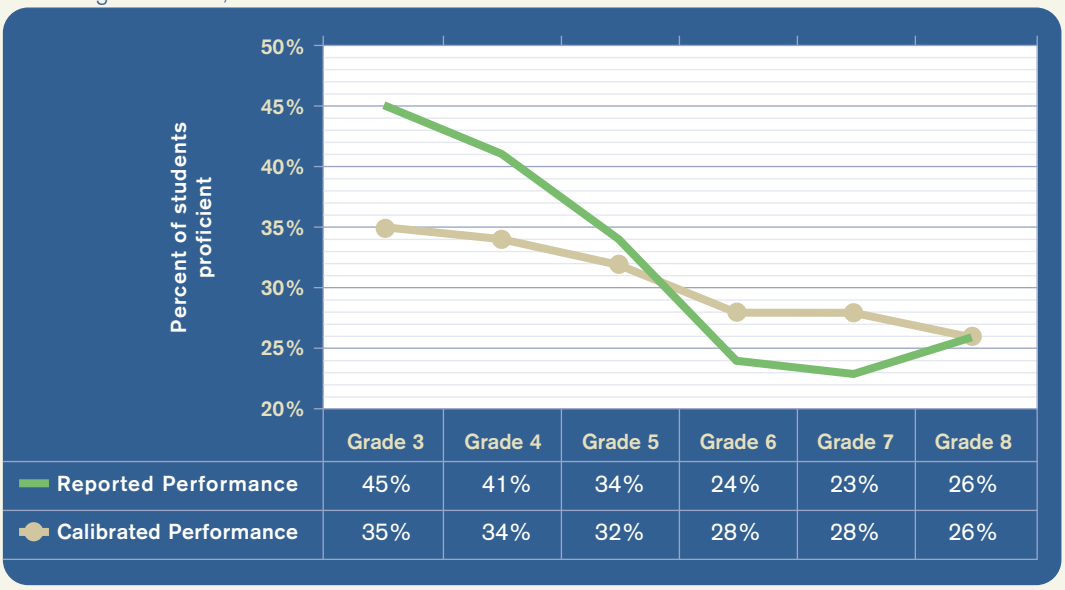
Figure 5 – New Mexico Reading Performance as Reported and as Calibrated to the Grade-Eight Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	55%	54%	57%	40%	50%	51%
Calibrated Performance	55%	53%	54%	50%	49%	51%

Note: This graphic shows, for example, that if New Mexico's grade-six reading cut score was set at the same level of difficulty as its grade-eight cut score, 50 percent of sixth graders would achieve the proficient level, rather than 40 percent, as was reported by the state.

Figure 6 – New Mexico Mathematics Performance as Reported and as Calibrated to the Grade-Eight Standard, 2006



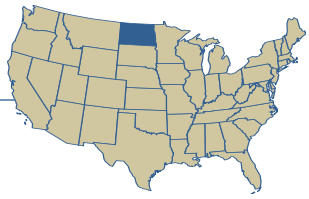
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	45%	41%	34%	24%	23%	26%
Calibrated Performance	35%	34%	32%	28%	28%	26%

Note: This graphic shows, for example, that if New Mexico's grade-three mathematics cut score was set at the same level of difficulty as its grade-eight cut score, 35 percent of third graders would achieve the proficient level, rather than 45 percent, as was reported by the state.

Policy Implications

New Mexico proficiency cut scores are relatively high in mathematics, at least compared to the other 25 states in this study. Its reading cut scores are about at the mid-point. This finding is fairly consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found New Mexico's standards to be in the upper-middle sector for reading and upper level for mathematics. Over the year-long span of time that cut scores were tracked for this study, the state's cut scores for mathematics have become less difficult in grades six and eight, although not in other grades. Nonetheless, New Mexico's expectations in mathematics are still not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. State policymakers might consider adjusting their math cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

North Dakota



Introduction

This study linked data from the 2004 and 2005 administrations of North Dakota's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that North Dakota's definitions of proficiency in reading and mathematics are generally consistent with the cut scores set by other 25 states in this study. In other words, North Dakota's tests are about average in terms of difficulty.

Yet the difficulty level of North Dakota's tests declined somewhat from 2004 to 2005—part of the No Child Left Behind Era—although not in all grades. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the North Dakota test not being matched by learning gains on the Northwest Evaluation Association test. One finding of this study is that North Dakota's proficiency cut scores are now relatively easier for third-grade students than for eighth graders, particularly in mathematics (taking into account the obvious differences in subject content and children's development). North Dakota policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: North Dakota State Assessment (NDSA)

North Dakota currently uses a fall assessment called the North Dakota State Assessment (NDSA), which tests reading/language arts and mathematics in grades 3 through 8 (the "NCLB grades"), and grade 11. Students are also tested for science in grades 4, 8, and 11. The current study analyzed reading and math results from a group of elementary and middle schools in which almost all students took both the state's assessment and MAP, using the fall 2004 and fall 2005 administrations of the two tests. (The methodology section of this report explains how performance on these two tests was compared.) These linked results were then used to estimate the scores on NWEA's scale that would be equivalent to the proficiency cut scores for each grade and subject on the North Dakota State Assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.)

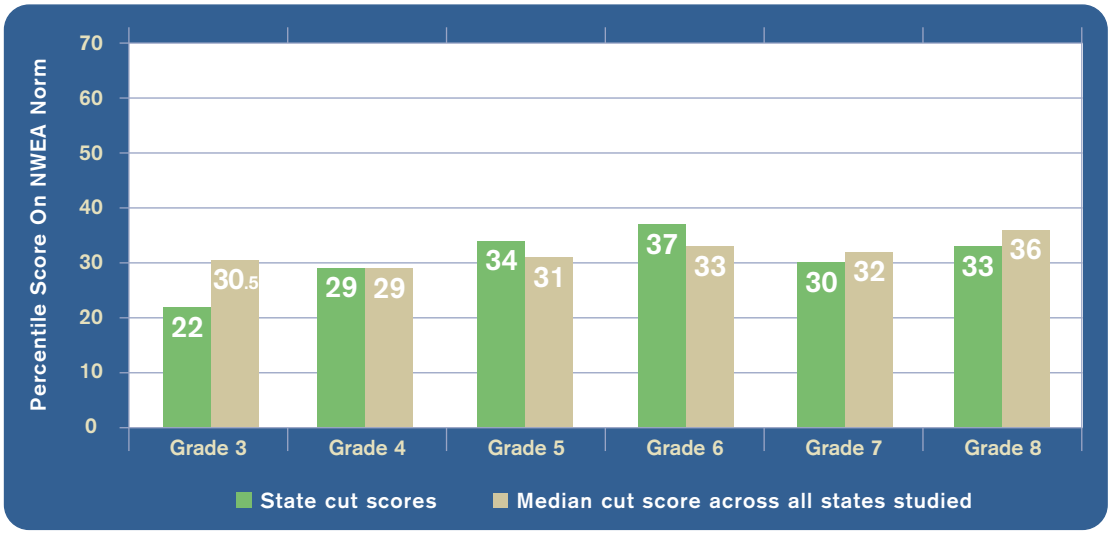
Part 1: How Difficult are North Dakota's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 percent would make it. How do we know a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of North Dakota's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the North Dakota cut score on a test of equivalent difficulty. The following two figures show the difficulty of North Dakota's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in North Dakota ranged between the 22nd and 37th percentiles, with the sixth grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 20th and 41st percentiles, with eighth grade being most challenging.

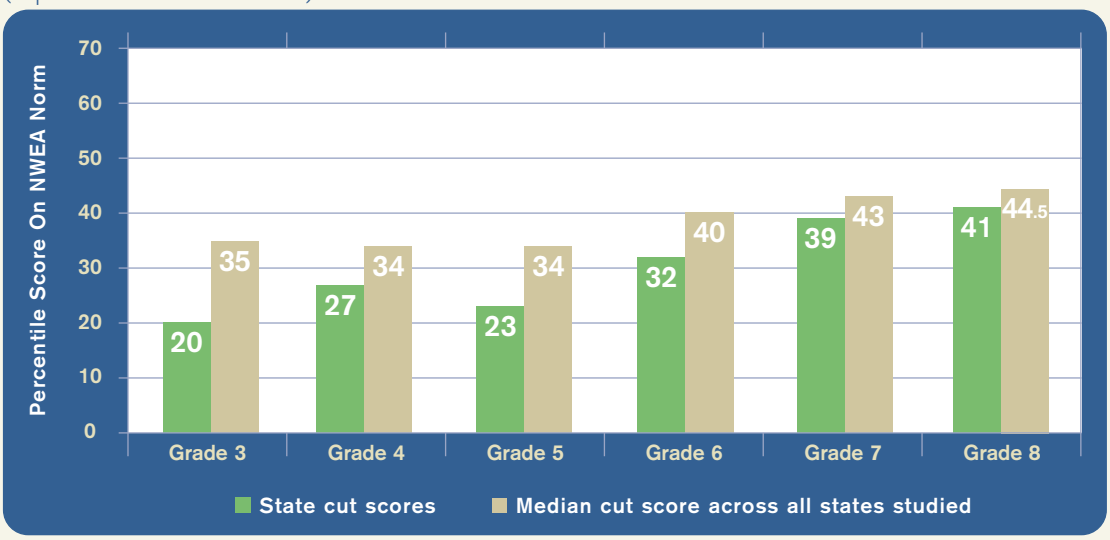
Another way of assessing difficulty is to evaluate how North Dakota's proficiency cut scores rank relative to other states in the study. Table 1 shows that the North Dakota cut scores generally rank in the lower half in difficulty among the 26 states studied for this report, and notably so in math. Its reading cut scores in grades 5 and 6 are its highest, ranking seventh and tenth, respectively.

Figure 1 – Estimate of North Dakota Reading Cut Scores in Relation to All 26 States Studied, 2005 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Only in grades 5 and 6 do North Dakota's cut scores surpass the median. The grade-3 cut score is particularly low.

Figure 2 – Estimate of North Dakota Mathematics Cut Scores in Relation to All 26 States Studied, 2005 (Expressed in MAP Percentiles)



Note: North Dakota's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. Across grades, North Dakota's math test cut scores are below the median, with differences ranging from 3.5 to 15 points.

Table 1 – North Dakota Rank for Proficiency Cut Scores Among States in Reading and Mathematics, 2005

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	20	13	7	10	18	14
Mathematics	21	20	22	19	17	13

Note: This table ranks North Dakota's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

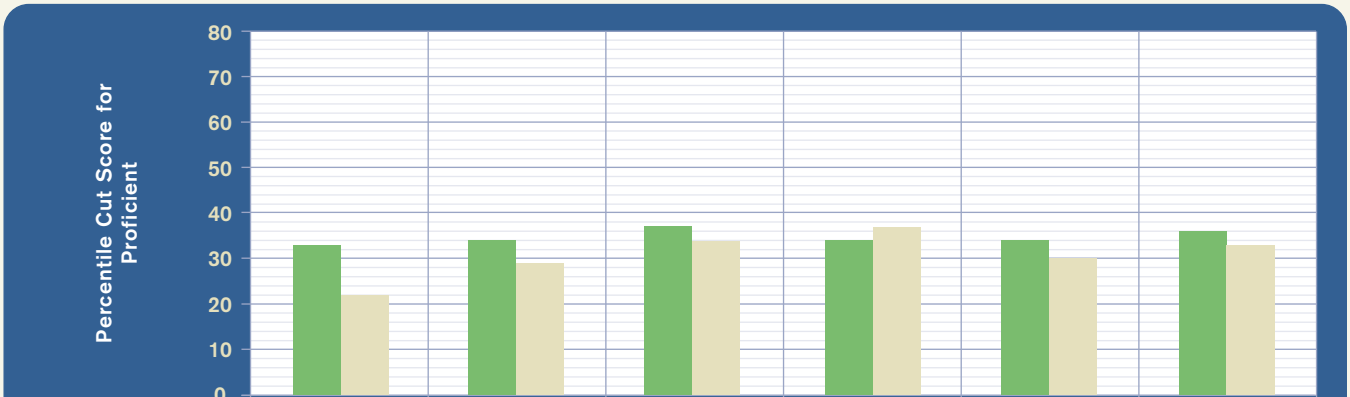
Part 2: Changes in Cut Scores over Time

In order to measure their consistency, North Dakota's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for both the 2004-05 and 2005-06 school years. Cut score estimates in both years were available in reading and mathematics for grades 3 through 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed.

Is it possible, then, to make comparisons of the proficiency scores between earlier administrations of North Dakota tests and today's? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height to judge proficiency. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measures or scales used by the NDSA in 2004 and in can be linked to the scale used to report MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the NDSA in 2004 and in 2005 on the MAP scale and ascertain whether the test may have changed in difficulty.

Figure 3 – Estimated Differences in North Dakota's Proficiency Cut Scores in Reading, 2004-2005 (Expressed in MAP Percentiles).



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Fall '04	33	34	37	34	34	36
Fall '05	22	29	34	37	30	33
Difference	-11	-5	-3	+3	-4	-3

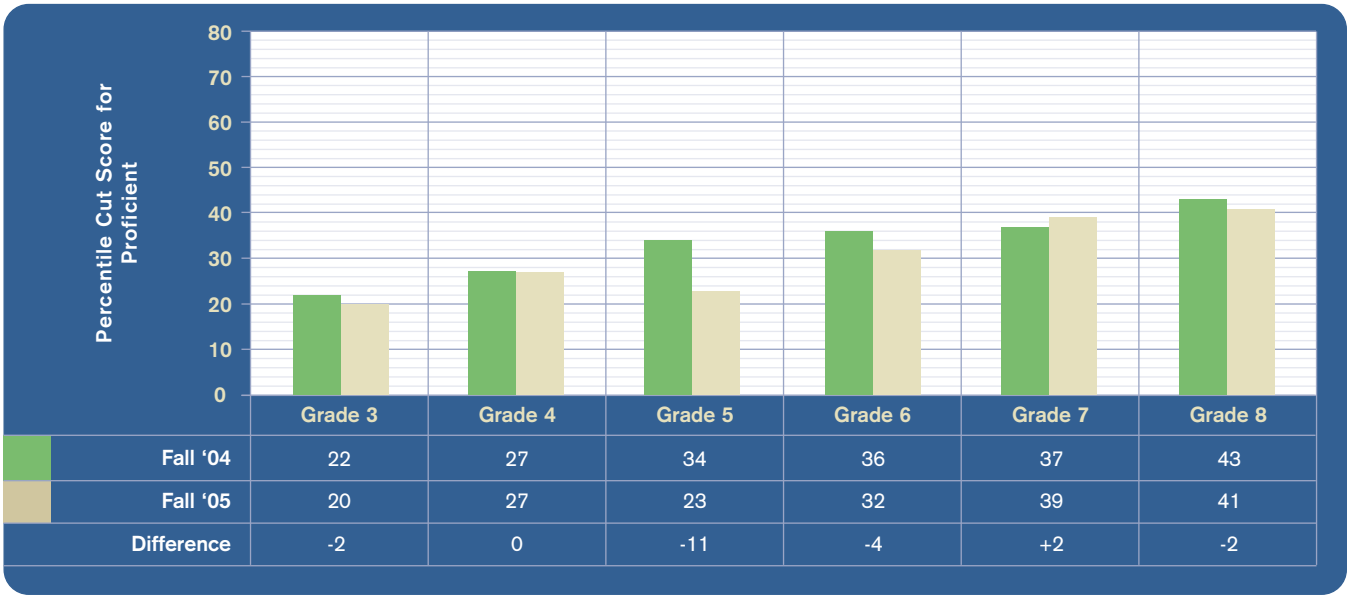
Note: This graphic shows how the difficulty of achieving proficiency reading has changed. For example, third-grade students in 2004 had to score at the 33rd percentile nationally in order to be considered proficient, while 2005 third graders only had to score at the 22nd percentile to achieve proficiency. The changes in all other grades were within the margin of error (in other words, too small to be considered substantive).

North Dakota's estimated **reading** analyses indicate a decrease in the third-grade cut score from 2004 to 2005 (see Figure 3), but no other substantive changes. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the third-grade reading proficiency rate in 2005 to be 11 percent higher than in 2004. (In fact, North Dakota reported no change in proficiency rating for third graders over this period.)

North Dakota's estimated **mathematics** cut scores showed a decrease in difficulty for fifth grade between the two years (Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, this would likely yield an 11 percent increase in the proficiency rate. (In fact, North Dakota reported no change in proficiency rate for fifth graders over this period.) No other substantive changes in math cut score cut scores were found.

Thus, one could fairly say that North Dakota's third-grade test in reading and fifth-grade test in mathematics were easier to pass in 2005 than in 2004, while the remaining tests were about the same.

Figure 4 – Estimated Differences in North Dakota's Proficiency Cut Scores in Mathematics, 2004-2005 (Expressed in MAP Percentile Ranks).



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Fall '04	22	27	34	36	37	43
Fall '05	20	27	23	32	39	41
Difference	-2	0	-11	-4	+2	-2

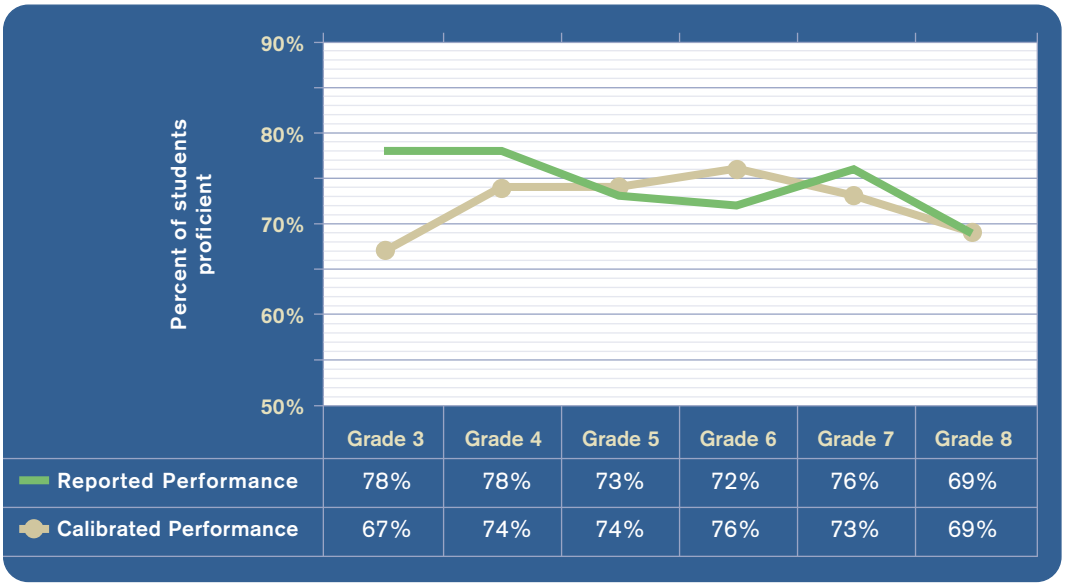
Note: This graphic shows how the difficulty of achieving proficiency has changed. For example, fifth-grade students in 2004 had to score at the 34th percentile nationally in order to be considered proficient, while in 2005 fifth graders had to score only at the 23rd percentile to achieve proficiency. The changes in all other grades were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth-grade cut score is no more or less difficult for eighth graders to achieve than the third-grade cut score is for third graders. When cut scores are all calibrated to the grade-eight standard, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the cut scores at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 showed that North Dakota's upper-grade cut scores in reading and mathematics were generally more challenging than in the lower grades, particularly for mathematics. (This was true for most states studied.) The two figures that follow show North Dakotans' reported performance on their state test in reading (Figure 5) and mathematics (Figure 6), compared with the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-eight standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades. This would lead to the conclusion that the higher rates of mathematics proficiency that the state has reported for younger students are somewhat misleading.

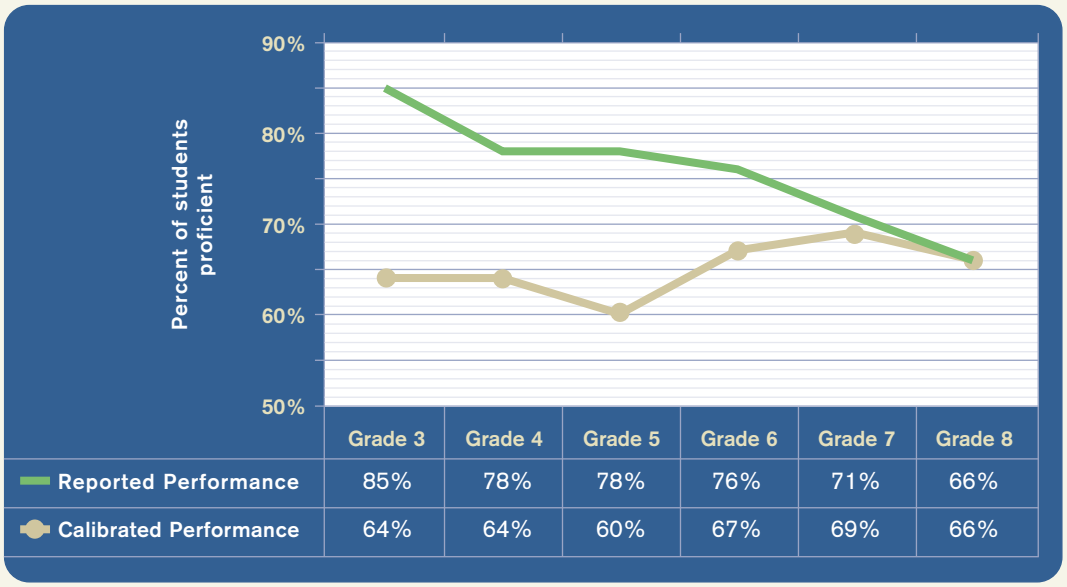
Figure 5 – North Dakota Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	78%	78%	73%	72%	76%	69%
Calibrated Performance	67%	74%	74%	76%	73%	69%

Note: This graphic shows, for example, that if North Dakota's grade-3 reading standard was set at the same level of difficulty as its grade-8 cut score, 67 percent of third graders would achieve the proficient level, rather than the 78 percent reported by the state.

Figure 6 – North Dakota Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



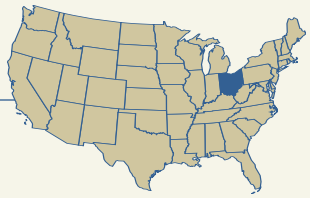
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	85%	78%	78%	76%	71%	66%
Calibrated Performance	64%	64%	60%	67%	69%	66%

Note: This graphic shows, for example, that if North Dakota's grade-3 mathematics cut score was set at the same level of difficulty as its grade-8 cut score, 64 percent of third graders would achieve the proficient level, rather than the 85 percent reported by the state.

Policy Implications

North Dakota's proficiency cut scores stand in the middle of the pack when compared to the other 25 states in this study. This finding is relatively consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which found North Dakota's standards to be in the upper-middle part of the distribution of all states studied. There appears to be a downward drift in some of the reading and mathematics cut scores, although not for all grades. Moreover, North Dakota's expectations are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. North Dakota policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Ohio



Introduction

This study linked data from the 2007* administration of Ohio's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that the difficulty of Ohio's proficiency cut scores in reading and math is generally below the median, compared to the 25 other states in the study.

Ohio's estimated reading cut scores are even in their difficulty across the grades studied, but its estimated mathematics cut scores are more difficult in the middle grades. As a result, reported proficiency rates for mathematics may not reflect true differences in performance across grades. State policy-makers might consider adjusting their math cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher, student, and school performance across these domains.

What We Studied: Ohio Achievement Tests (OAT)

Ohio currently uses an assessment called the Ohio Achievement Tests (OAT), which assess mathematics and reading in grades 3-8. The current study linked reading and math data from spring 2007 administrations to a common scale also administered in the 2007 school year.

To determine the difficulty of Ohio's proficiency cut scores, we linked data from Ohio's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Ohio's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high jump bar is easy to leap? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high jump bar is challenging? We know because only one (or perhaps none) of those same 100 individuals would successfully meet that level of challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

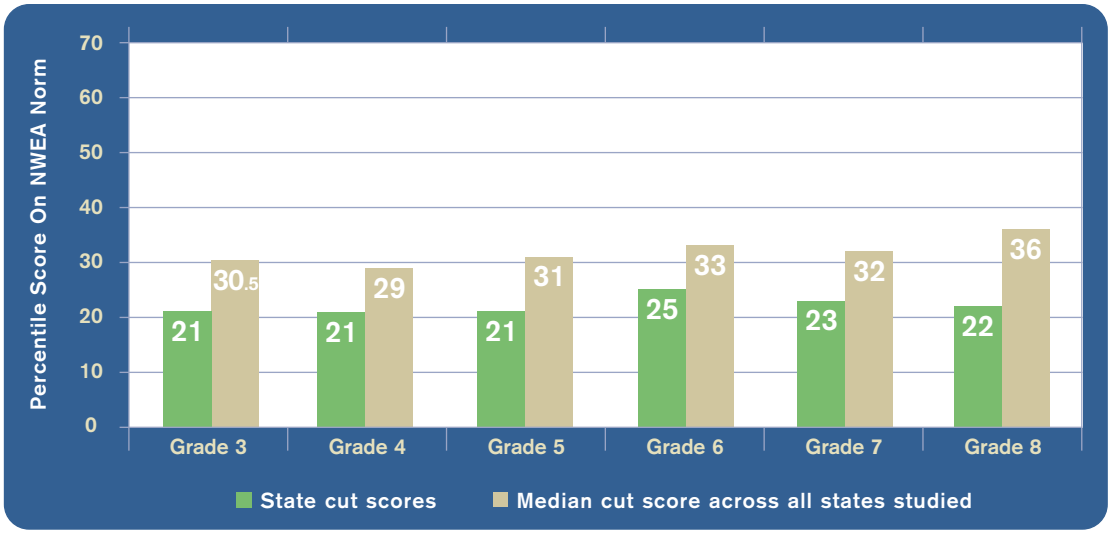
*The Ohio report uses data collected from the 2007 testing season, rather than the 2006 season as with most other state reports, since the distribution of schools comprising the 2007 sample represented a better cross-section of the state than were available for the 2006 sample.

Applying the concept to this assignment, we evaluated the difficulty of the Ohio proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the cut score on a test of equivalent difficulty. The following two figures show the estimated difficulty of Ohio's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2007 in relation to the median cut score for all the states in the study, and compared to the NWEA norm group. The estimated proficiency cut scores for **reading** in Ohio ranged between the 21st and 25th percentiles on NWEA norms, with the sixth grade cut score being most challenging. In **mathematics**, the estimated cut scores ranged between the 20th and 40th percentiles, with fifth grade being most challenging.

Ohio's estimated reading cut scores in every grade are below the median level of difficulty among the states studied. Estimated mathematics cut scores are also below the median in all but grade five. Note, too, that Ohio's reading cut scores are lower than its math cut scores in every grade beyond the third. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Ohio students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

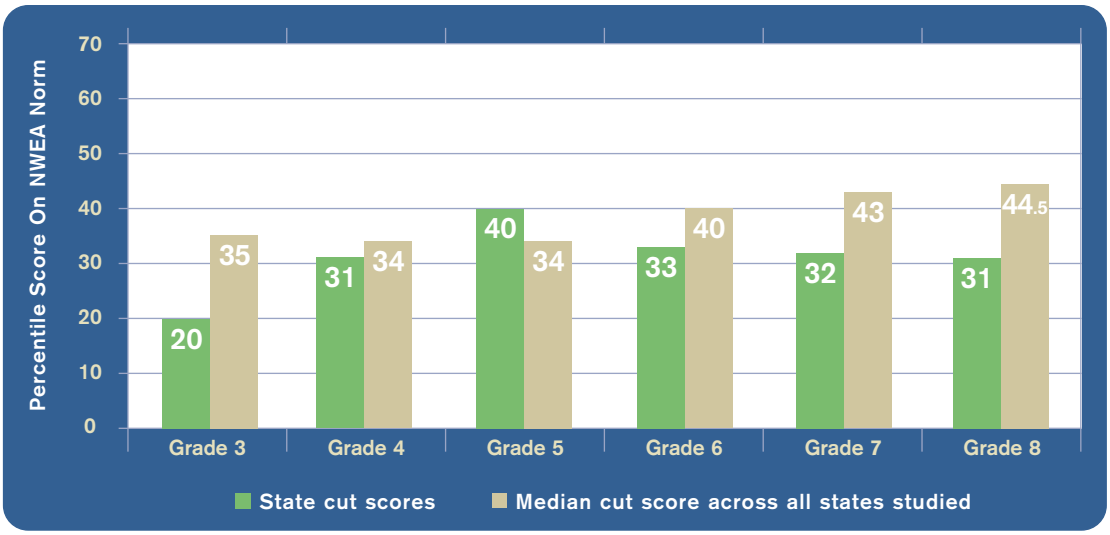
Another way of assessing difficulty is to evaluate how Ohio's proficiency cut scores rank relative to other states. Table 1 shows that Ohio's estimated reading and mathematics cut scores generally rank among the lower half of the 26 states examined for this report.

Figure 1 – Ohio Reading Cut Scores in Relation in Relation to All 26 States Studied, 2007 (Expressed in MAP Percentiles)



Note: This figure compares estimated reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Across all grades, Ohio's reading scores are below the median, with differences ranging from 8 to 14 points.

Figure 2 – Ohio Mathematics Cut Scores in Relation to All 26 States Studied, 2007 (Expressed in MAP Percentiles).



Note: Ohio's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. Only in grade 5 do Ohio's standards surpass the median. In grades 3, 7, and 8, the state's cut scores are well below the median.

Table 1 – Ohio Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2007

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	21	22	23	20	22	21
Mathematics	20	17	9	17	21	19

Note: This table ranks Ohio's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

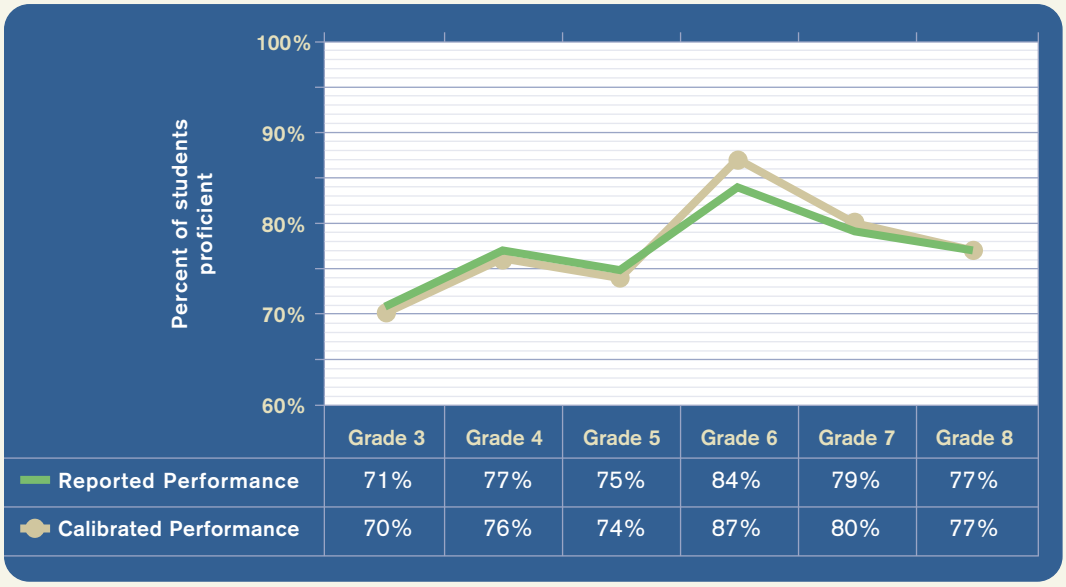
Part 2: Calibration across Grades*

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth grade cut score is no more or less difficult for eighth graders to achieve than the third grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third grade proficiency cut score puts a student on track to eventually achieve the cut scores in eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 showed the relative difficulty levels of the reading and mathematics cut scores, illustrating the fluctuation across grades. Those figures showed that the difficulty of the estimated cut scores was very stable across the grades in reading, but that the mathematics cut scores started out easy, peaked in grade five, then eased up a bit. The following two figures show Ohio's reported performance in reading (Figure 3) and mathematics (Figure 4) on the state test, compared with the proficiency rates that would be achieved if the cut scores were all calibrated to the grade 8 standard. Because the estimated reading cut scores are so well calibrated to begin with, Figure 3 shows very little difference between reported proficiency rates and what those rates would like if they were calibrated to the grade 8 cut score. Figure 4, however, shows that the reported proficiency rates in mathematics may actually be overestimating the percentage of third grade students who are actually on track to meet the eighth proficiency standards.

* Ohio was one of seven states in this study for which cut score estimates could be determined for only one year. Therefore, it was not possible to examine whether its cut scores have changed over time.

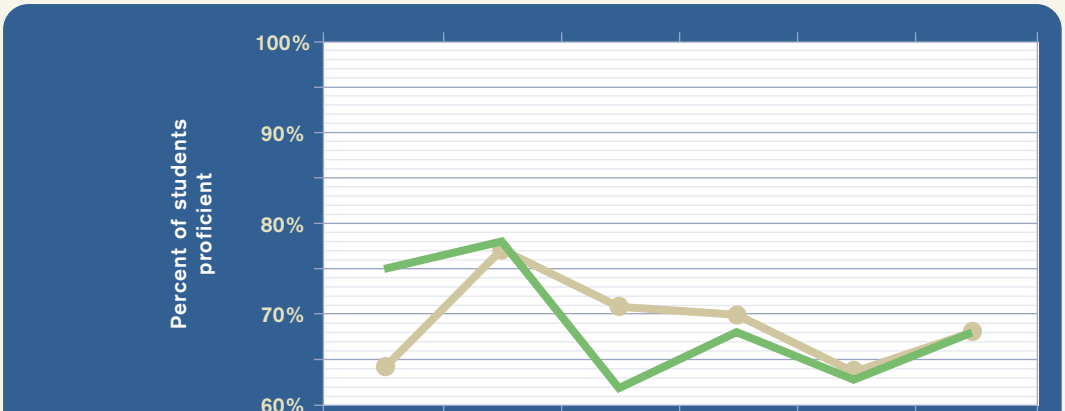
Figure 3 – Ohio Reading Performance as Reported and as Calibrated to the Grade 8 Standard, 2007



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	71%	77%	75%	84%	79%	77%
Calibrated Performance	70%	76%	74%	87%	80%	77%

Note: This graphic shows, for example, that if Ohio's grade-three reading cut score was set at the same level of difficulty as its grade-eight cut score, 75 percent of third graders would achieve the proficient level, rather than 71 percent, as was reported by the state.

Figure 4 – Ohio Mathematics Performance as Reported and as Calibrated to the Grade 8 Standard, 2007



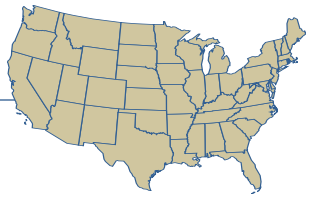
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	75%	77%	62%	68%	63%	68%
Calibrated Performance	64%	77%	71%	70%	64%	68%

Note: This graphic shows, for example, that if Ohio's grade-3 mathematics cut score were as difficult as its grade-8 cut score, 64 percent of third graders would achieve the proficient level, rather than 75 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what constitutes proficiency, Ohio is a bit below the median in both reading and mathematics, at least compared to the other 25 states in this study. Ohio's proficiency cut scores are well calibrated from grade to grade in reading, but less so for mathematics. As a result, reported mathematics proficiency rates may slightly exaggerate differences across grades. State policymakers might consider adjusting the difficulty of their math cut scores across grades so that parents and schools can be assured that proficient performance at the earlier grades accurately predicts proficiency at the later grades. Furthermore, state leaders need to be aware of the disparity between math and reading standards when evaluating differences in teacher and student performance across these domains.

Rhode Island



Introduction

This study linked data from the 2005 administration of Rhode Island's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Rhode Island's definitions of proficiency in reading and mathematics are relatively consistent with the standards set by the other 25 states in this study, with its reading tests a bit above average in difficulty and its math tests a bit below average.

In addition, we found Rhode Island's cut scores to be less challenging for third-grade students than for eighth graders. State policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: New England Common Assessment Program (NECAP)

Rhode Island currently uses a fall assessment called the New England Common Assessment Program (NECAP), developed in conjunction with New Hampshire and Vermont. NECAP tests students in grades three through eight in English/language arts and mathematics. Science tests and standards are currently under development. The current study uses linked reading and math data from the fall 2005 NECAP administration (in New Hampshire schools, which use the same assessment tool and proficiency cut scores) to a common scale also administered during the 2005-6 school year.

To determine the difficulty of Rhode Island's proficiency cut scores, we linked reading and math data from Rhode Island's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Rhode Island's Definitions of Proficiency in Reading and Math?

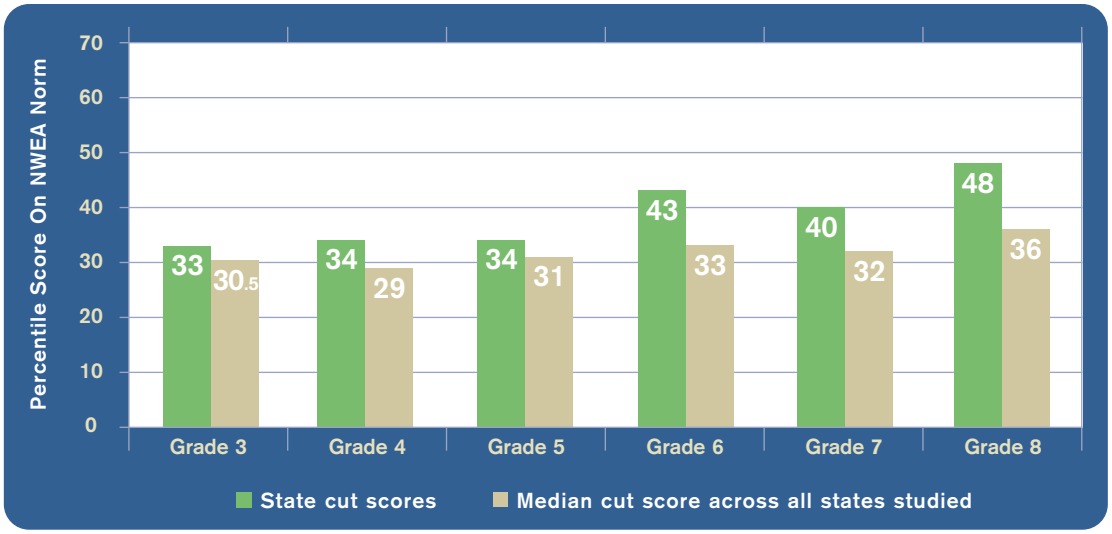
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of Rhode Island's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Rhode Island cut score on a test of equivalent difficulty. The following two figures show the difficulty of Rhode Island's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Rhode Island ranged between the 33rd and 48th percentiles for the norm group, with the eighth-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 34th and 53rd percentiles, with eighth grade again being most challenging.

Rhode Island's cut scores in both reading and mathematics are consistently at or above the median in difficulty among the states studied. Note, though, that Rhode Island's cut scores for reading are generally lower than its cut scores for mathematics at the same grade. (This was the case in the majority of states studied.) Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Rhode Island students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

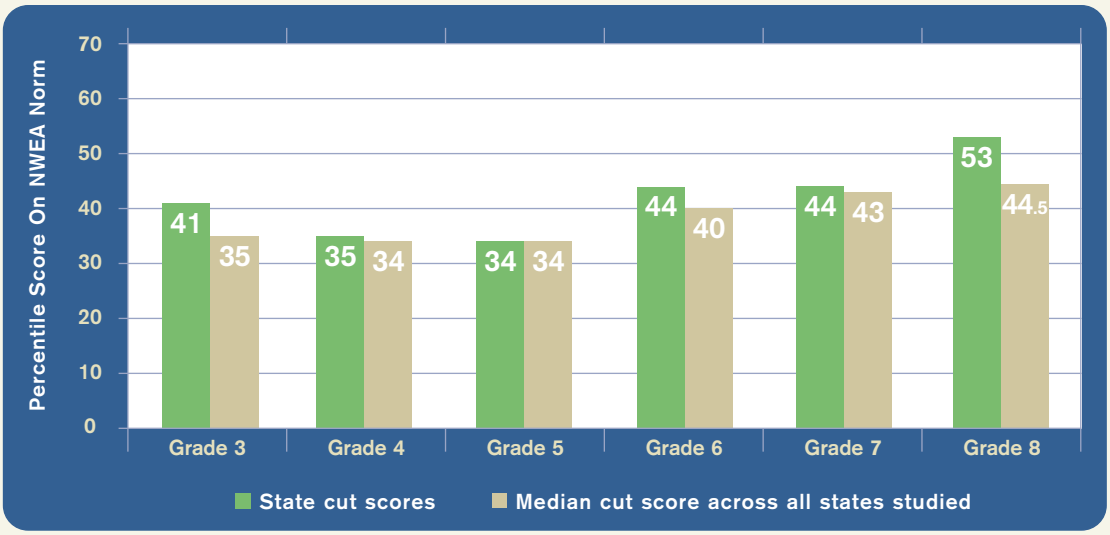
Another way of assessing difficulty is to evaluate how Rhode Island's proficiency cut scores rank relative to other states. Table 1 shows that Rhode Island's cut scores generally rank in the upper third for reading and at about the middle for math among the 26 states studied for this report. Its reading cut score in grade eight is particularly high, ranking third out of 26 states.

Figure 1 – Rhode Island Reading Cut Scores in Relation to All 26 States Studied, 2005 (expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Rhode Island's cut scores are consistently 2.5 to 12 percentiles above the median.

Figure 2 – Rhode Island Mathematics Cut Scores in Relation to All 26 States Studied, 2005 (expressed in MAP Percentiles)



Note: Rhode Island's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. The cut scores are consistently 1 to 8.5 percentiles above the median, except in grade five, where the cut score is precisely equal to the median.

Table 1 – Rhode Island Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2005

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	9	6	7	4	7	3
Mathematics	8	10	13	9	9	6

Note: This table ranks Rhode Island's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

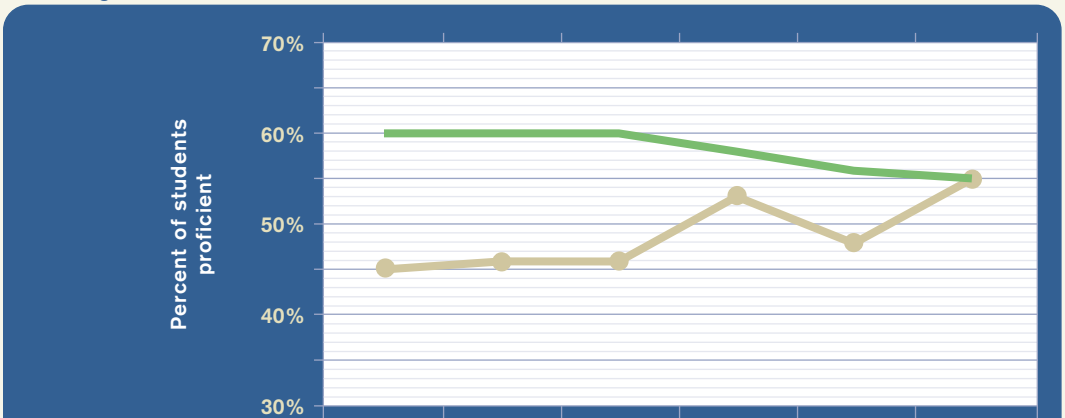
Part 2: Calibration across Grades*

Calibrated proficiency cut scores are relatively equal in difficulty across all grades. Thus, the eighth-grade cut score is no more or less difficult for eighth graders to achieve than the third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

* Rhode Island was one of seven states in this study for which cut score estimates could be determined for only one year. Therefore, it was not possible to examine whether its cut scores have changed over time.

Figures 1 and 2 showed the relative difficulty of the reading and mathematics cut scores across the different grades, indicating that that the upper-grade cut scores in reading and mathematics were somewhat more challenging than the cut scores in the lower grades. (This was the case for the majority of states studied.) The following two figures show Rhode Island's reported performance in reading (Figure 3) and mathematics (Figure 4) on its state test and the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-eight standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades. This would lead to the conclusion that the stronger rates of proficiency that the state has reported for lower grades students are somewhat misleading.

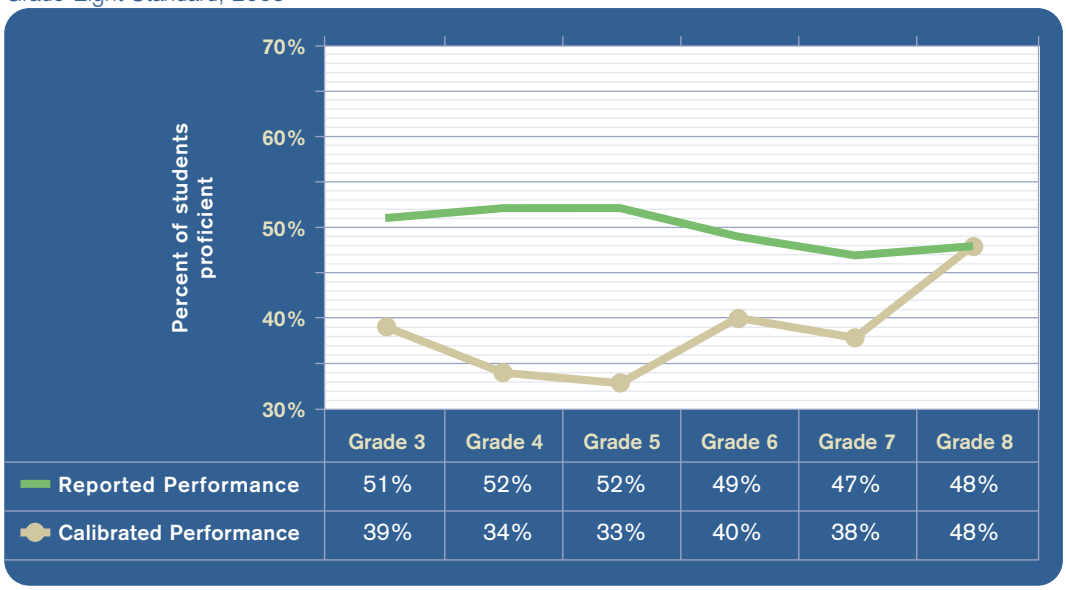
Figure 3 – Rhode Island Reading Performance as Reported and as Calibrated to the Grade-Eight Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	60%	60%	60%	58%	56%	55%
Calibrated Performance	45%	46%	46%	53%	48%	55%

Note: This graphic shows, for example, that if Rhode Island's grade-3 reading cut score was set at the same level of difficulty as its grade-8 cut score, 45 percent of third graders would achieve the proficient level, rather than 60 percent, as was reported by the state.

Figure 4 – Rhode Island Mathematics Performance as Reported and as Calibrated to the Grade-Eight Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	51%	52%	52%	49%	47%	48%
Calibrated Performance	39%	34%	33%	40%	38%	48%

Note: This graphic shows, for example, that if Rhode Island's grade-3 mathematics cut score was set at the same level of difficulty as its grade-8 cut score, 39 percent of third graders would achieve the proficient level, rather than 51 percent, as was reported by the state.

Policy Implications

When determining what constitutes proficiency in reading and math, Rhode Island is about in the middle of the pack, at least compared to the other 25 states in this study. It's noteworthy that Rhode Island's cut scores are not smoothly calibrated across grades, though. Students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. State policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

South Carolina

Introduction

This study linked data from the 2002 and 2006 administrations of South Carolina's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that South Carolina's definitions of proficiency in reading and mathematics are relatively difficult, compared to the cut scores set by the 25 other states in the study. In other words, South Carolina's tests are well above average in terms of difficulty.

Yet the difficulty level of South Carolina tests' decreased somewhat from 2002 to 2006—the No Child Left Behind era—and quite dramatically in a few grades. South Carolina's current reading test is easier in third, fourth, and fifth grades than it was in 2002, as is the math test for sixth and eighth grades. There are many possible explanations for these declines (see pp. 34-35 of the main report), which were caused by learning gains on the South Carolina test not being matched by learning gains on the Northwest Evaluation Association test. One finding of this study is that South Carolina's reading cut scores are relatively easier in the early grades than they are for eighth graders (taking into account the differences in subject content and children's development). State policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: South Carolina Palmetto Achievement Challenge Tests (PACT)

South Carolina currently uses an assessment called the South Carolina Palmetto Achievement Challenge Tests (PACT), which tests mathematics, English/language arts, science, and social studies in grades 3 through 8. The same set of tests was used in spring 2002 to test students in mathematics and English/language arts in grades 3 through 8. The current study linked reading and math results from spring 2002 and spring 2006 administrations in a group of elementary and middle schools to a common scale also administered in the 2002 and 2006 school years.

To determine the difficulty of South Carolina's proficiency cut scores, we linked data from South Carolina's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of schools in which almost all students had taken both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are South Carolina's Definitions of Proficiency in Reading and Math?

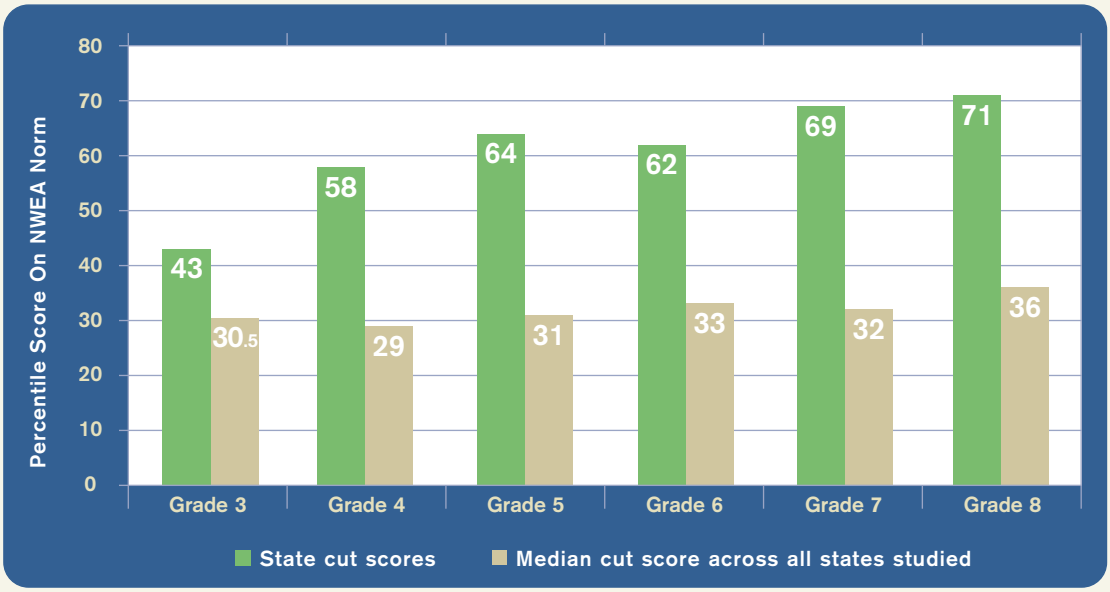
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of South Carolina's proficiency standards by estimating the proportion of students in NWEA's norm group who would perform above the South Carolina standard on a test of equivalent difficulty. The following two figures show the difficulty of South Carolina's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in South Carolina ranged between the 43rd and 71st percentiles nationally, with the eighth grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 64th and 75th percentiles, with eighth grade again the most challenging.

Across grades 3 through 8, South Carolina's cut scores in both reading and mathematics are consistently more difficult than the median cut scores of the other states in the study, and above the performance of the average student of that grade within the NWEA norm group. Note, though, that South Carolina's cut scores for reading are generally lower than for mathematics. (This pattern was spotted in the majority of states studied.) Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, South Carolina students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentages that pass state tests in those subjects.

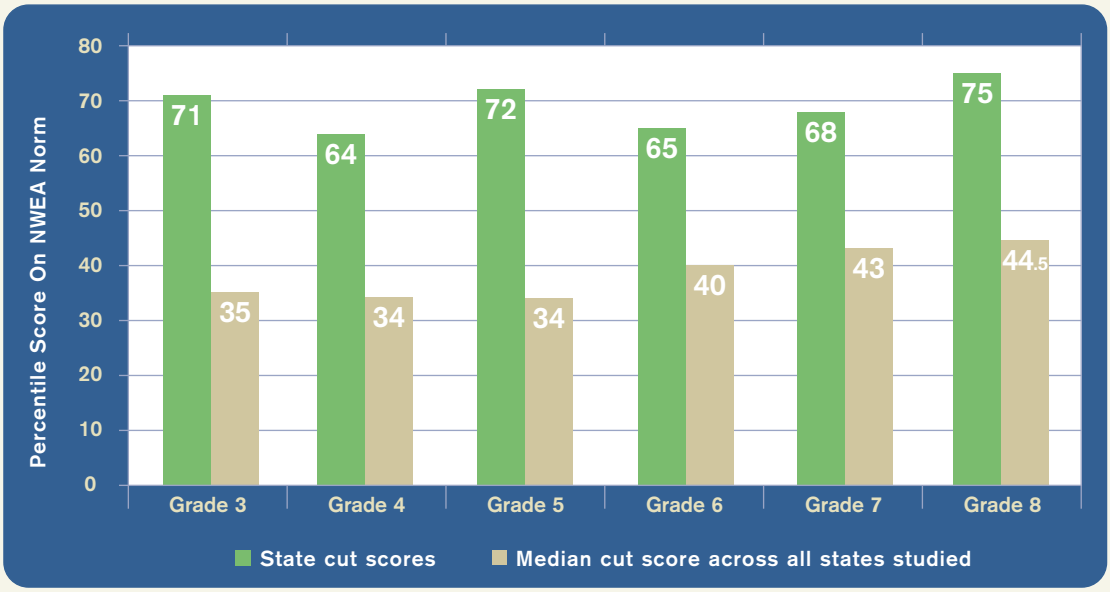
Another way of assessing difficulty is to evaluate how South Carolina's proficiency cut scores rank relative to other states in the study. Table 1 shows that the South Carolina cut scores generally rank among the very top of the 26 states studied for this report.

Figure 1 – South Carolina Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. South Carolina's cut scores across all grades are above the median, ranging from 12.5 to 37 percentile points above.

Figure 2 – South Carolina Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: South Carolina's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. Across all grades, the state's cut scores surpass the median by 25 to 38 points.

Table 1 – South Carolina Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	4	2	1	1	1	1
Mathematics	1	2	1	2	2	1

Note: This table ranks South Carolina's cut scores relative to the cut scores of the other 25 states in the study. South Carolina ranks number one in four grades for reading and in three grades for mathematics.

Part 2: Changes in Cut Scores over Time

In order to measure their consistency, South Carolina's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2001-2 and 2005-6 school years. Cut score information for reading and mathematics were available for both years in grades three through eight.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Plus, unintentional drift can occur even in states, such as South Carolina, that maintained their proficiency levels.

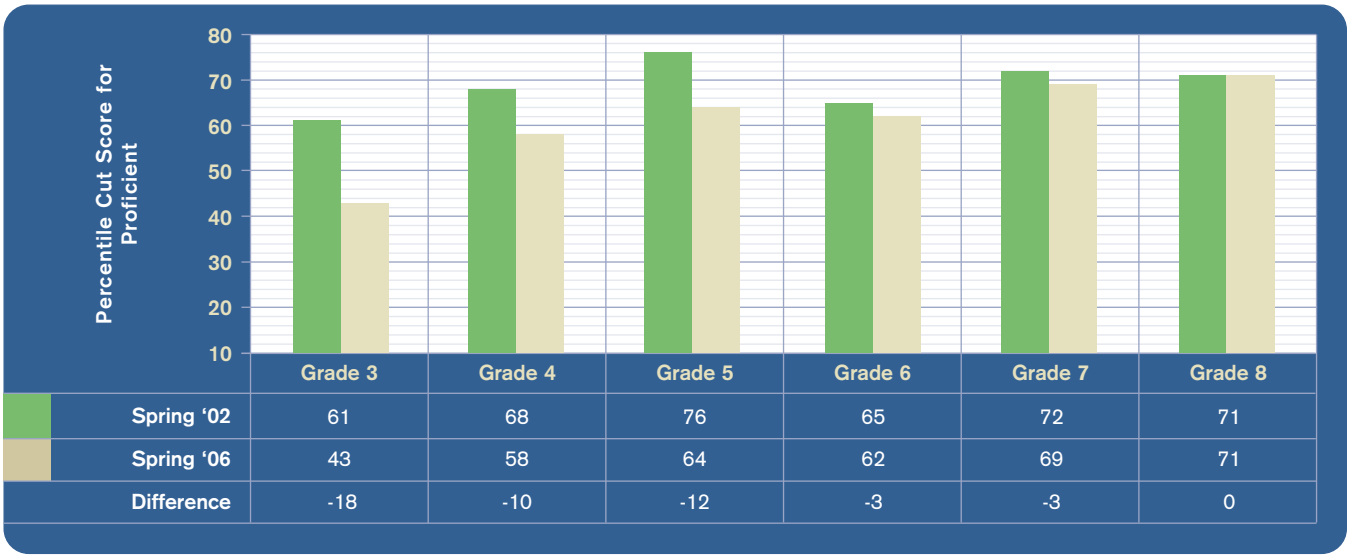
Is it possible, then, to compare the proficiency scores across a four-year period? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measures or scales used by the PACT in 2002 and in 2006 can both be linked to the scale that was used to report MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the PACT in 2002 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty. This allows us to estimate whether the PACT in 2006 was easier or harder than in 2002.

South Carolina's estimated **reading** cut scores (see Figure 3) decreased over this four-year period for third, fourth, and fifth grades, with no substantial changes in proficiency cut scores at the higher grades. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the third-, fourth-, and fifth-grade reading proficiency rates in 2006 to be 18 percent, 10 percent, and 12 percent higher, respectively, than in 2002. (South Carolina reported a 13-point gain for third graders, an 8-point gain for fourth graders, and a 9-point gain for fifth graders over this period.)

South Carolina's estimated **mathematics** cut scores (see Figure 4) showed substantive decreases for grades 6 and 8, with all other grades' cut scores remaining essentially the same. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect 7 and 5 percent increases in the mathematics proficiency rates reported in 2006 for sixth- and eighth-grade pupils, respectively. (South Carolina reported an 8-point gain for sixth graders and a 3-point gain for eighth graders over this period.)

Thus, one could fairly say that South Carolina's reading tests were easier to pass in 2006 than they were in 2002 for the lower grades, but about the same for the higher grades. Similarly, the math tests were easier to pass in grades 6 and 8, but about the same in the other grades. As a result, any increased proficiency rates reported for grades in which the cut scores grew easier may not be entirely a product of improved student achievement.

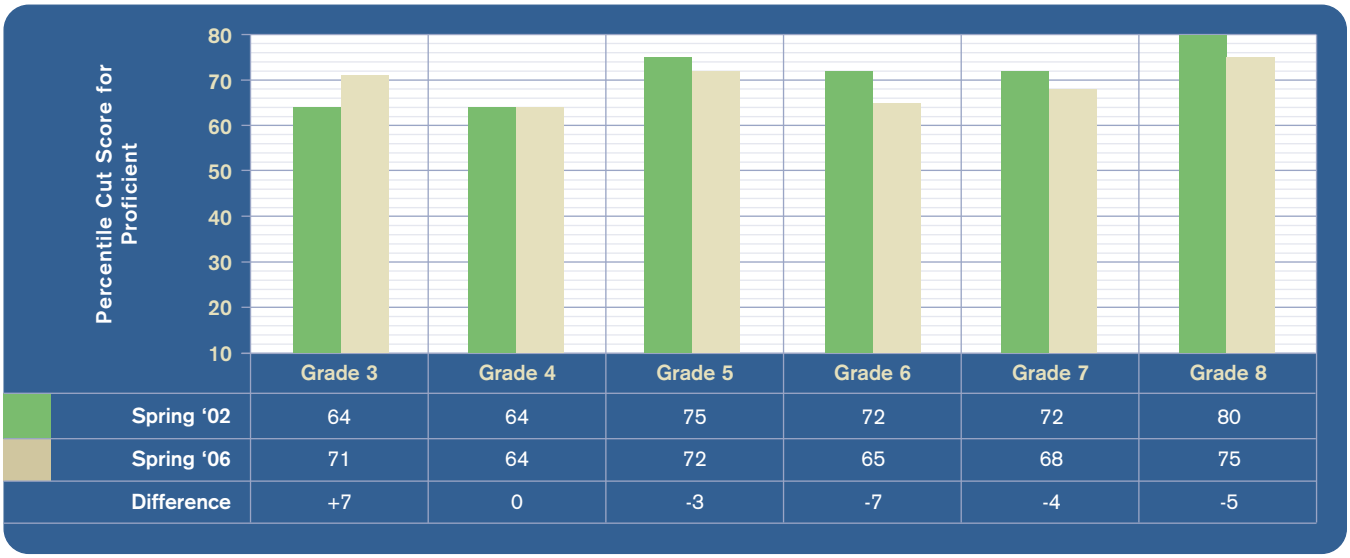
Figure 3 – Estimated Differences in South Carolina's Proficiency Cut Scores in Reading, 2002-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Spring '02	61	68	76	65	72	71
Spring '06	43	58	64	62	69	71
Difference	-18	-10	-12	-3	-3	0

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, third-grade students in 2002 had to score at the 61st percentile of the NWEA norm nationally in order to be considered proficient, while in 2006 third graders had to score at the 43rd percentile of the NWEA norm to achieve proficiency. The changes in grades 6, 7, and 8 were within the margin of error (in other words, too small to be considered substantive).

Figure 4 – Estimated Change in South Carolina's Proficiency Cut Scores in Mathematics, 2002-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Spring '02	64	64	75	72	72	80
Spring '06	71	64	72	65	68	75
Difference	+7	0	-3	-7	-4	-5

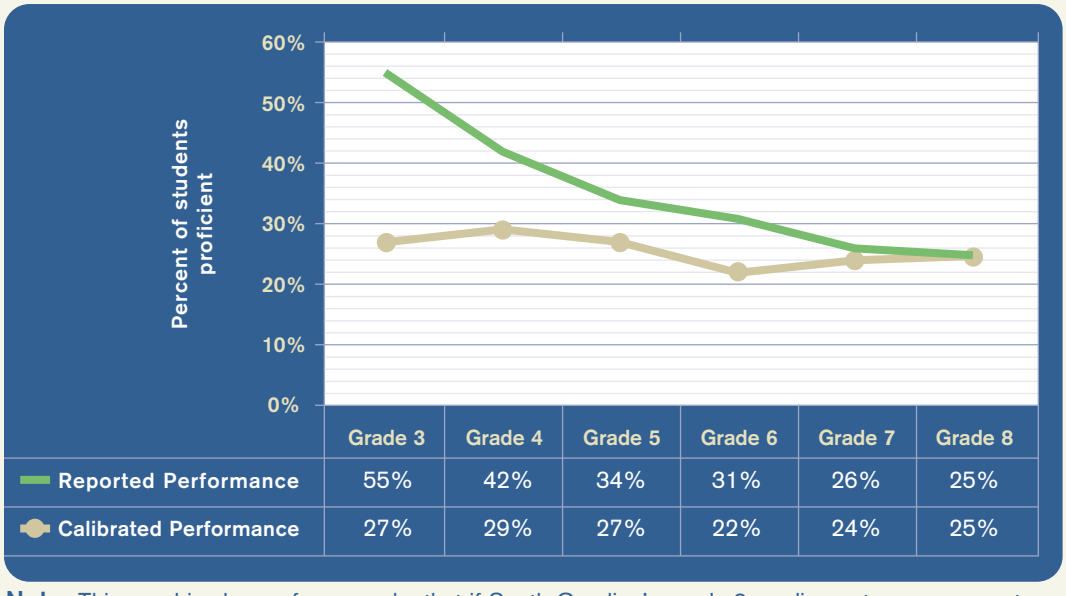
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, sixth-grade students in 2002 had to score at the 72nd percentile of the NWEA norm group in order to be considered proficient, while in 2006 sixth graders only had to score at the 65th percentile of the NWEA norm to achieve proficiency. The changes in grades 3, 4, 5, and 7 were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 showed that South Carolina's upper-grade cut scores in reading in 2006 were considerably more challenging than in the lower grades, while the mathematics cut scores were fairly well calibrated. The two figures that follow show South Carolina's reported performance in reading (Figure 5) and mathematics (Figure 6) on the state test compared with the proficiency rates that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut scores are removed, student performance is more consistent at all grades. This would lead to the conclusion that the higher rates of reading proficiency that the state has reported for lower-grade students are somewhat misleading. Specifically, the apparent decline across grades may be an artifact of differences in the difficulty of the cut scores, and not because of differences in actual student performance.

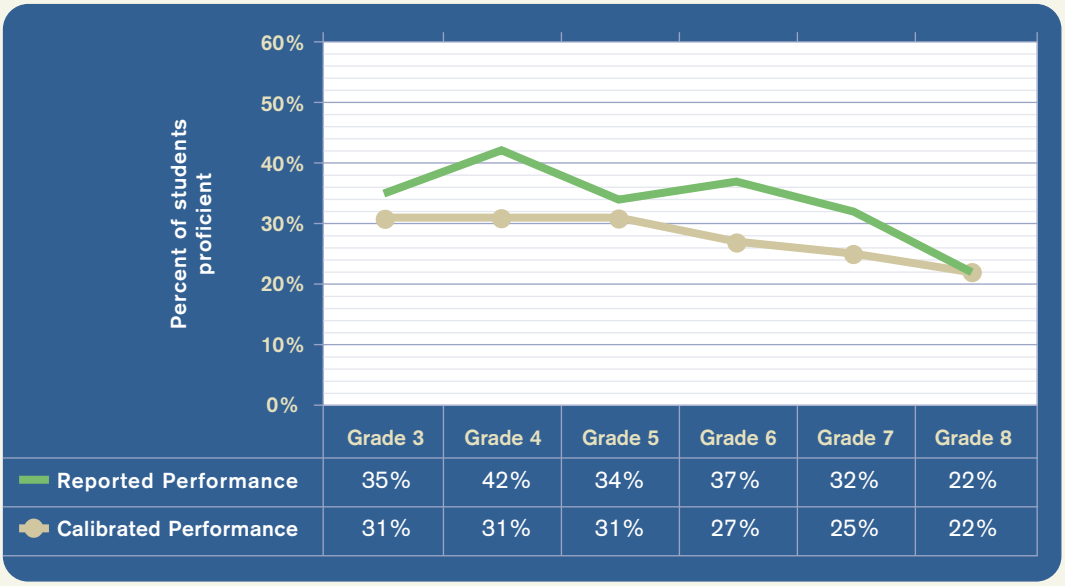
Figure 5 – South Carolina Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	55%	42%	34%	31%	26%	25%
Calibrated Performance	27%	29%	27%	22%	24%	25%

Note: This graphic shows, for example, that if South Carolina's grade-3 reading cut score was set at the same level of difficulty as its grade-8 cut score, 27 percent of third graders would achieve the proficient level, rather than 55 percent, as was reported by the state.

Figure 6 – South Carolina Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



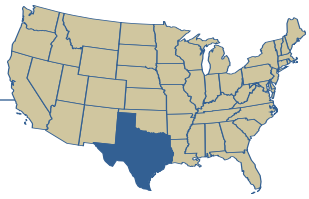
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	35%	42%	34%	37%	32%	22%
Calibrated Performance	31%	31%	31%	27%	25%	22%

Note: This graphic shows, for example, that if South Carolina's grade-3 mathematics standard was set at the same level of difficulty as its grade-8 cut score, 31 percent of third graders would achieve the proficient level, rather than 35 percent, as was reported by the state.

Policy Implications

South Carolina's proficiency cut scores in reading and math are relatively high, at least compared with the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found South Carolina's standards to be among the highest in the country. In the past several years, however, the difficulty of these cut scores has declined, though not in all grades. As a result, South Carolina's expectations are not smoothly calibrated across grades, at least in reading; students who are proficient in third grade are not necessarily on track to be proficient by the eighth grade. South Carolina policy-makers might consider adjusting their reading cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Texas



Introduction

This study linked data from the 2003 and 2006 administrations of Texas's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Texas's definitions of proficiency are relatively less difficult than the cut scores set by the other 25 states in this study in reading and mathematics. In other words, Texas's tests are below average in terms of difficulty.

Still, the level of difficulty has increased from 2003 to 2006—the No Child Left Behind era—though more so for some grades than others. Texas is one of the few states in this study whose cut scores have become more challenging over time. Even so, the state's expectations are not consistent from one grade to the next and policymakers should consider more closely calibrating them to ensure equivalent difficulty at all grades. In this way, parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Texas Assessment of Knowledge and Skills (TAKS)

Texas currently uses the Texas Assessment of Knowledge and Skills (TAKS), which tests students in reading in grades 3 through 9; in writing in grades 4 and 7; in English/language arts in grades 10 and 11; in mathematics in grades 3 through 11; in science in grades 5, 10, and 11; and social studies in grades 8, 10, and 11. The Spanish TAKS is administered in grades 3 through 6. Satisfactory performance on the TAKS at grade 11 is prerequisite to a high school diploma. TAKS was first administered in the 2002-2003 school year.

To determine the difficulty of Texas's proficiency cut scores, we linked data from state reading and math tests from a group of elementary and middle schools to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Texas's Definitions of Proficiency in Reading and Math?

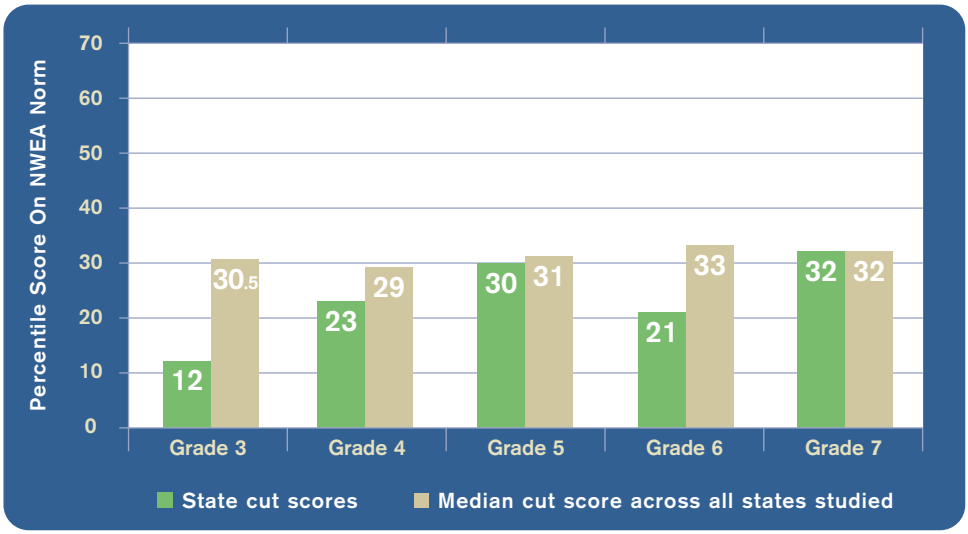
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of Texas's proficiency standards by estimating the proportion of students in NWEA's norm group who would perform above the Texas standard on a test of equivalent difficulty. The following two figures show the difficulty of Texas's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. Sample sizes were sufficient to generate cut score estimates for reading and math in grades 3 through 7. Grade-8 cut scores were not available. The proficiency cut scores for **reading** in Texas ranged between the 12th and 32nd percentiles nationally, with the seventh grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 24th and 41st percentiles with the seventh grade again being most challenging.

For most grade levels, Texas's cut scores in both reading and mathematics are below the median level of difficulty among the states studied. Note, though, that Texas's cut scores for reading are generally less difficult than the corresponding mathematics cut scores within a given grade. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Texas students may be performing worse in reading and better in mathematics than is apparent by looking at the percentage of students passing state tests in those subjects.

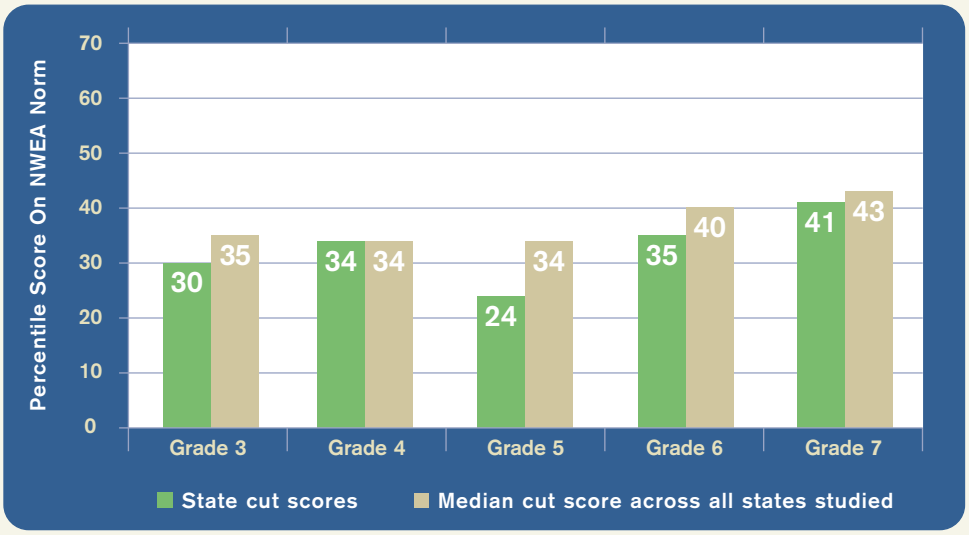
Another way of assessing difficulty is to evaluate how Texas's proficiency cut scores rank relative to other states. Table 1 shows that the Texas cut scores generally rank in the lower half for reading and the upper half for mathematics, among the 26 states studied for this report. Texas's third- and fourth-grade reading cut scores are particularly low, besting only two and six other states in the study, respectively. On the other hand, Texas ranks relatively high in third- and fourth-grade math.

Figure 1 – Estimate of Texas Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Only in grades 5 and 7 do Texas's cut scores approach or equal the median.

Figure 2 – Estimate of Texas Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: Texas's math-test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. Only in fourth grade does Texas's cut score reach the median.

Table 1 – Texas Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)				
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Reading	24	20	14	22	13
Mathematics	14	13	20	16	15

Note: This table ranks Texas's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Differences in Cut Scores over Time

In order to measure their consistency, Texas's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2002-3 and 2005-6 school years. Cut score estimates for both years were available for grades 3 through 7 for reading and grades 4 and 7 for mathematics.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed.

This was certainly the case for Texas. When the Texas Assessment of Knowledge and Skills (TAKS) was introduced in 2002-03, the Texas Education Agency formally adopted cut scores that would increase in difficulty over the first three years of testing. This was meant to give schools and students an opportunity to adjust to the new test and its expectations.

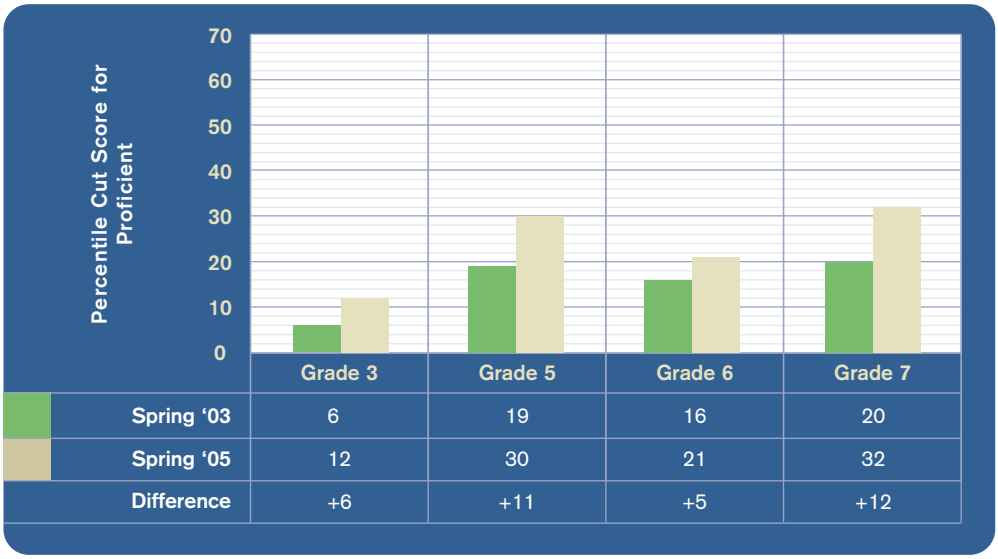
Is it possible, then, to compare the proficiency scores across this three-year period? Yes. Assume that we're judging a group of fourth graders on their high-jump prowess and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measures or scales used by the TAKS in 2003 and 2006 can both be linked to the scale that was used to report MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the TAKS in 2003 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty.

Texas's estimated **reading** cut scores indicate that, as intended by the state, the proficiency cut scores increased in difficulty over this three-year period for all available grades (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the reading proficiency rates in 2006 to be lower than they were in 2003. These more difficult cut scores would likely yield 6 percent, 11 percent, 5 percent, and 12 percent decreases in the proficiency rates for third, fifth, sixth, and seventh grade students, respectively. (Texas reported an 8-point decline for grade 7, although proficiency rates in grades 3, 5 and 6 actually increased by 4, 1, and 5 points, respectively.)

Texas's estimated **mathematics** cut scores showed similar patterns, with increases over three years in the difficulty of the proficiency cut scores for grades 5 and 7 (see Figure 4). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, these higher proficiency cut scores would likely yield decreases of 11 percent and 16 percent in the math proficiency rates for fifth and seventh graders, respectively. (Texas reported a 5-point decline for fifth graders and a 3-point decline for seventh graders over this period.)

Thus, one could fairly say that Texas's tests were harder to pass in 2006 than in 2003. As a result, improvements in actual student performance were been masked somewhat by the increased difficulty of the state's proficiency cut scores.

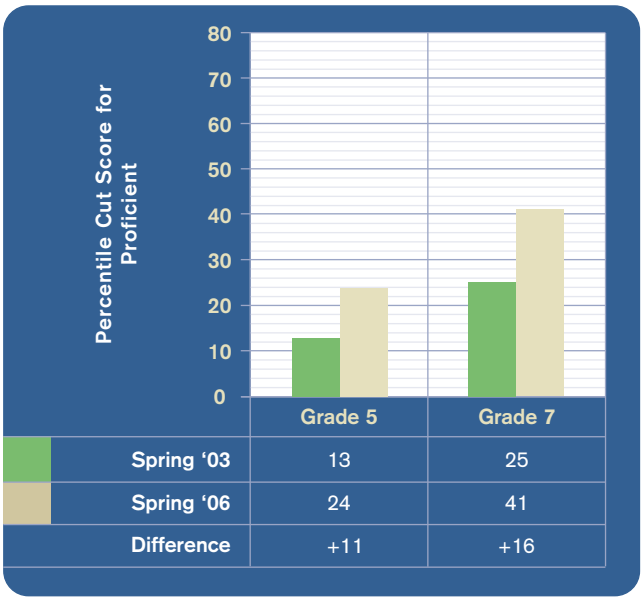
Figure 3 – Estimated Differences in Texas's Proficiency Cut Scores in Reading, 2003-2006 (Expressed in MAP Percentiles)



	Grade 3	Grade 5	Grade 6	Grade 7
Spring '03	6	19	16	20
Spring '05	12	30	21	32
Difference	+6	+11	+5	+12

Note: This graphic shows how the degree of difficulty in achieving proficiency in reading has changed. For example, third-grade students in 2003 had to score at the 6th percentile on the NWEA norm group in order to be considered proficient, while in 2006 third graders had to score at the 12th percentile of the NWEA norm group to achieve proficiency.

Figure 4 – Estimated Differences in Texas's Proficiency Cut Scores in Mathematics, 2003-2006 (Expressed in MAP Percentiles)



	Grade 5	Grade 7
Spring '03	13	25
Spring '06	24	41
Difference	+11	+16

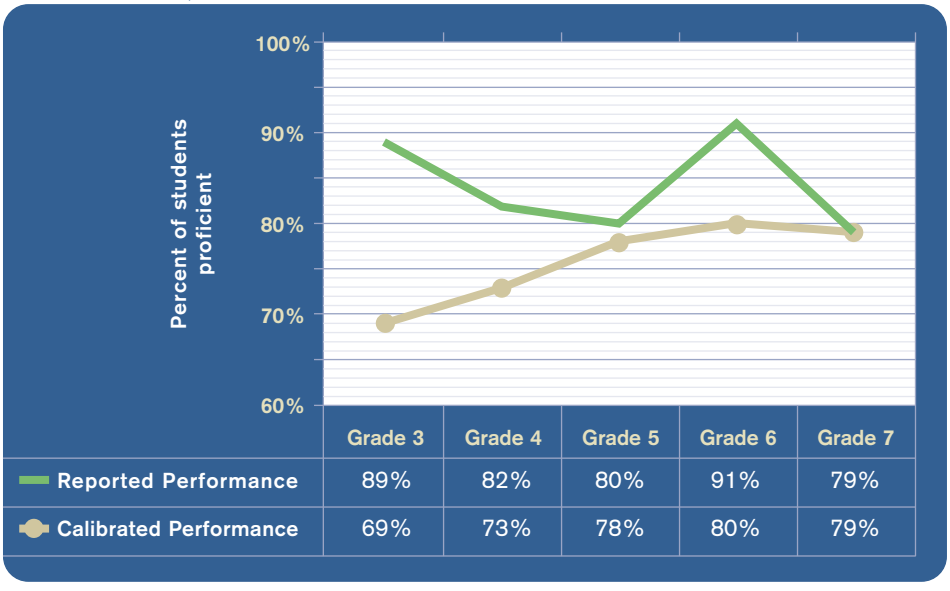
Note: This graphic shows how the degree of difficulty in achieving proficiency in math has changed. For example, fifth-grade students in 2003 had to score at the 13th percentile on the NWEA norm group in order to be considered proficient, while in 2006 fifth graders had to score at the 24th percentile of the NWEA norm group to achieve proficiency.

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Figures 1 and 2 showed that Texas's upper-grade cut scores in reading and mathematics were more challenging than the cut scores in the lower grades, particularly in grade 3. The two figures that follow show Texas's reported performance in reading (Figure 5) and mathematics (Figure 6) on the state test compared with the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-7 standard. When differences in grade-to-grade difficulty of the cut score are removed, student performance is more consistent at all grades. This would lead to the conclusion that the higher rates of proficiency that the state has reported for elementary school students are somewhat misleading.

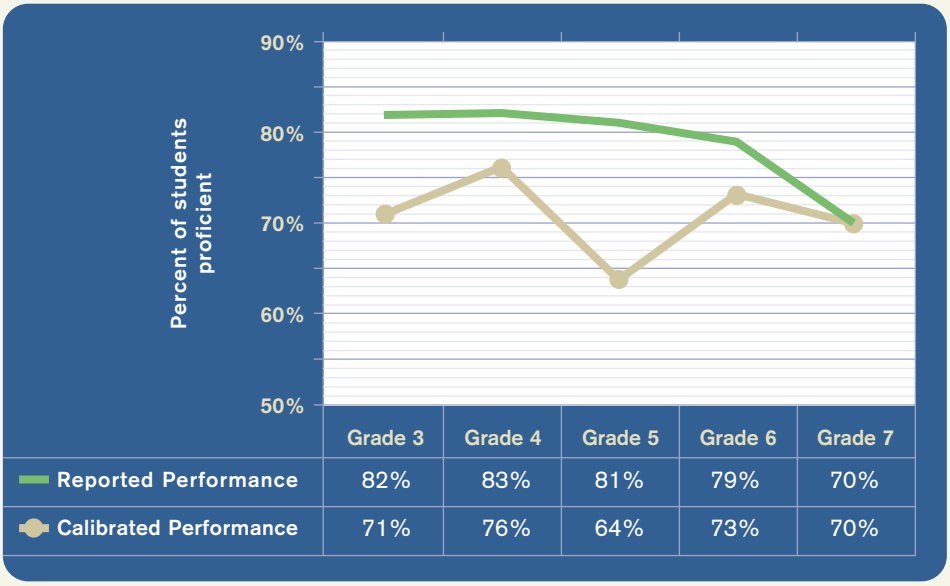
Figure 5 – Texas Reading Performance as Reported and as Calibrated to the Grade-7 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Reported Performance	89%	82%	80%	91%	79%
Calibrated Performance	69%	73%	78%	80%	79%

Note: This graphic shows, for example, that if Texas's grade-3 reading cut score was set at the same level of difficulty as its grade-7 cut score, 69 percent of third graders would achieve the proficient level, rather than 89 percent, as was reported by the state.

Figure 6 – Texas Mathematics Performance as Reported and as Calibrated to the Grade-7 Standard, 2006



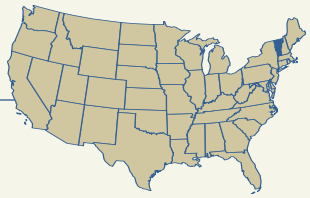
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
Reported Performance	82%	83%	81%	79%	70%
Calibrated Performance	71%	76%	64%	73%	70%

Note: This graphic shows, for example, that if Texas's grade-3 mathematics cut score was set at the same level of difficulty as its grade-7 cut score, 71 percent of third graders would achieve the proficient level, rather than 82 percent, as was reported by the state.

Policy Implications

When determining what constitutes proficiency, Texas is relatively low—more so in reading than in math—compared with the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Texas's reading standards to be in the bottom third of the distribution of all 50 states, and the mathematics standards closer to the middle. In recent years, the difficulty of the proficiency cut scores has increased, though some grades have increased more than others. As a result, Texas's expectations are not smoothly calibrated across grades; students who are proficient in third grade are not necessarily on track to be proficient by the seventh grade. Texas policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Vermont



Introduction

This study linked data from the fall 2005 administration of Vermont's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Vermont's definitions of proficiency in reading and mathematics are relatively consistent with the standards set by the other 25 states in this study, with its reading tests a bit above average in difficulty and its math tests a bit below average.

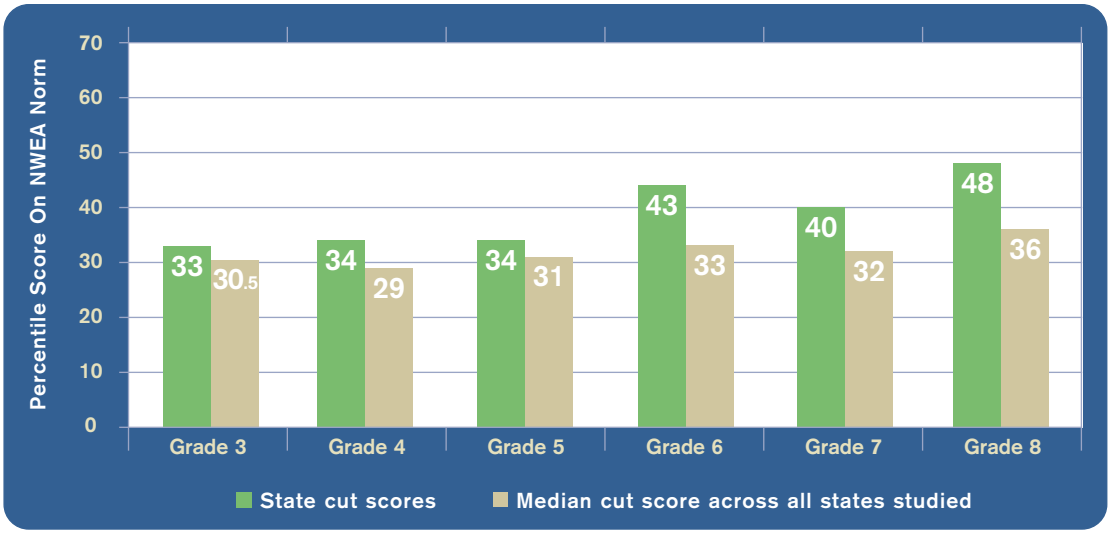
We also found Vermont's cut scores to be less challenging for third-grade students than for eighth graders. Vermont policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: New England Common Assessment Program (NECAP)

Vermont currently uses a fall assessment called the New England Common Assessment Program (NECAP), developed in conjunction with New Hampshire and Rhode Island. NECAP tests students in grades 3 through 8 in English/language arts and mathematics, with science tests and standards currently under development. The current study uses linked reading and math data from the fall 2005 NECAP administration (in New Hampshire schools, which use the same assessment tool and proficiency cut score standards) to a common scale also administered during the 2005-6 school year.

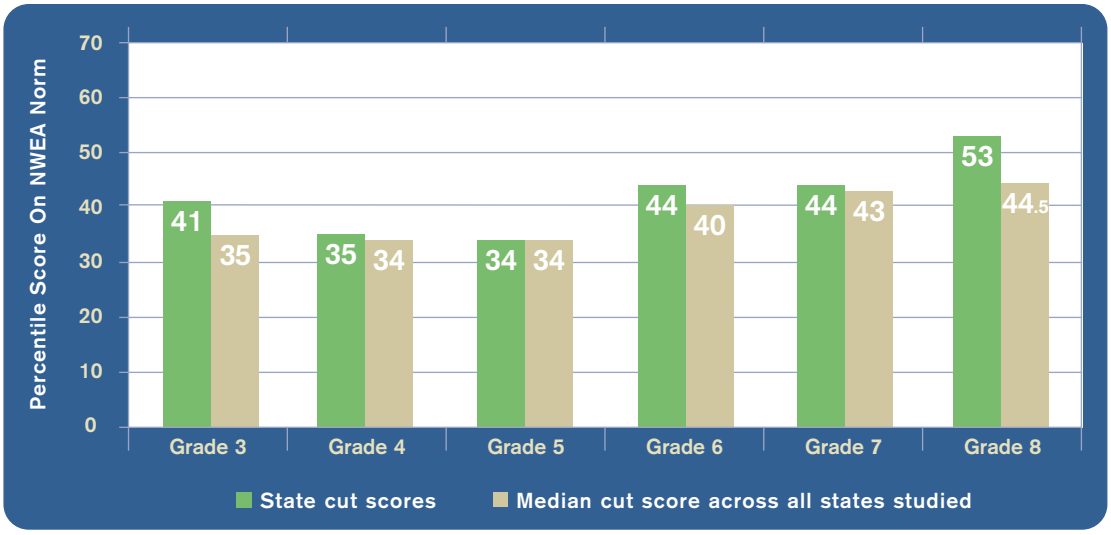
To determine the difficulty of Vermont's proficiency cut scores, we linked reading and math data from Vermont's tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Figure 1 – Vermont Reading Cut Scores in Relation to All 26 States Studied, 2005
(as Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Vermont's cut scores are consistently 2.5 to 12 percentile points above the median.

Figure 2 – Vermont Mathematics Cut Scores in Relation to All 26 States Studied, 2005
(as Expressed in MAP Percentiles)



Note: Vermont's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut score of all 26 states reviewed in this study. The cut scores are consistently 1 to 8.5 percentile points above the median, with the exception of grade 5 where the state's cut score is at the median.

Part 1: How Difficult are Vermont's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this assignment, we evaluated the difficulty of Vermont's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Vermont cut score on a test of equivalent difficulty. The following two figures show the difficulty of Vermont's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Vermont ranged between the 33rd and 48th percentiles for the norm group, with the eighth grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 34th and 53rd percentiles, with eighth grade again being the most challenging.

Vermont's cut scores in both reading and mathematics are consistently at or above the median in difficulty among the states studied. Note, though, that Vermont's cut scores for reading are generally lower than for math at the same grades. (This was the case in the majority of states studied.) Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Vermont students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentages passing state tests in those subjects.

Another way of assessing difficulty is to evaluate how Vermont's proficiency cut scores rank relative to other states. Table 1 shows that the Vermont cut scores generally rank in the upper third for reading and at about the middle for math among the 26 states studied for this report. Its reading cut score in grade 8 is particularly high, ranking third out of the 26 states.

Table 1 – Vermont Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2005

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	9	6	7	4	7	3
Mathematics	8	10	13	9	9	6

Note: This table ranks Vermont's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Calibration across Grades*

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

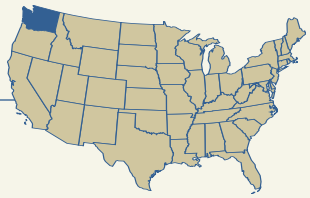
Examining Vermont's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed the relative difficulty of Vermont's reading and mathematics cut scores across the different grades, indicating that that the upper-grade cut scores in both subjects were somewhat more challenging than in the lower grades. (This was the case for the majority of states studied.) In other states within the current study, it was possible to show how these differences in cross-grade difficulty affect the proficiency rates (the percentages of students reported as "proficient" or better within each grade), and what the proficiency rates would look like if the cut scores were all calibrated to the eighth-grade difficulty level. Unlike other states, however, Vermont's State Department of Education website does not publish its proficiency rate data by grade, so such analyses were not possible. In other states with patterns of difficulty similar to Vermont's Figures 1 and 2, however, we saw that differences in proficiency rates, and in particular, dips in performance at the middle-school grades, typically were minimized when the difficulty of the cut scores were standardized. Such patterns suggested that dips in performance in middle-school grades were at least in part the product of non-calibrated cut scores rather than real differences in student performance across grades.

*Vermont was one of seven states in this study for which cut score estimates could be determined for only one year. Therefore, it was not possible to examine whether its cut scores have changed over time.

Policy Implications

When determining what constitutes proficiency in reading and math, Vermont was about in middle of the pack, at least compared to the other 25 states in this study. Vermont's cut scores are not smoothly calibrated across grades, however, which makes it difficult for the public to accurately evaluate observed differences in student performance across grades. State policymakers might consider adjusting their cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Washington



Introduction

This study linked data from the 2004 and 2006 administrations of Washington's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Washington's definitions of proficiency in reading and mathematics are relatively challenging in comparison to the standards set by the other 25 states in this study. In other words, Washington's tests are above average in terms of difficulty.

The level of difficulty stayed about the same from 2004 to 2006—during the No Child Left Behind era—except for fourth-grade reading, where it became easier.

This study found that Washington's mathematics cut scores are relatively easier for the earlier grades than for the higher grades (taking into account the differences in subject content and children's development). State policymakers might consider adjusting Washington's cut scores to ensure equivalent difficulty at all grades so that elementary school students are on track to be proficient in the later grades.

What We Studied: Washington Assessment of Student Learning (WASL)

Washington currently uses a spring assessment called the Washington Assessment of Student Learning (WASL), which tests reading and math in grades 3 through 8 and grade 10, as required by NCLB. Students are also tested in science in grades 5, 8, and 10, and in writing in grades 4, 7 and 10. The current study linked reading and math data from the spring 2004 and spring 2006 WASL administrations to a common scale also administered in the 2004 and 2006 school years.

To determine the difficulty of Washington's proficiency cut scores, we linked data from state tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Washington's Definitions of Proficiency in Reading and Math

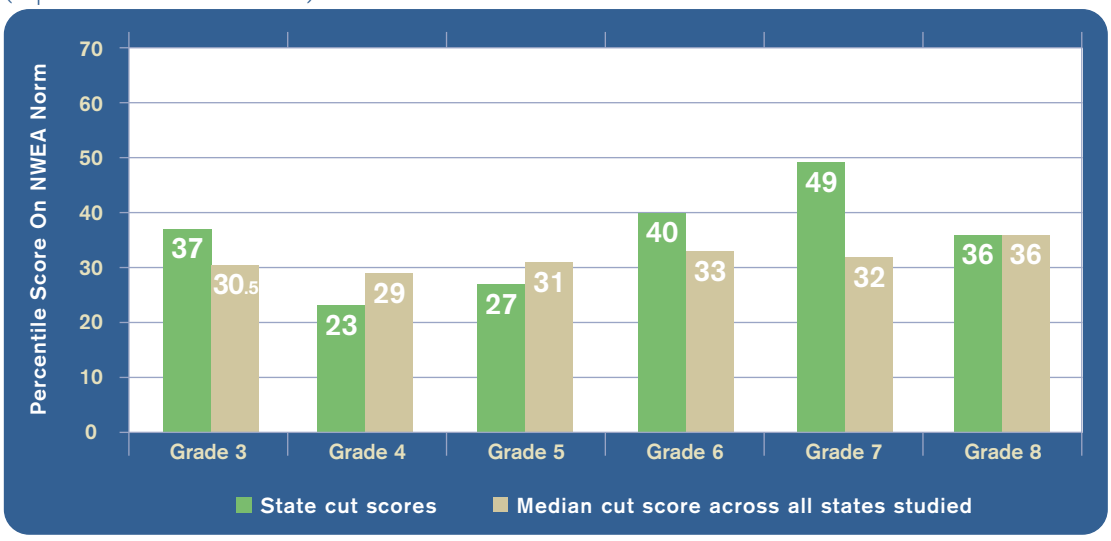
One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot-high bar is easy to jump over? We know because, if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

Applying that approach to this task, we evaluated the difficulty of Washington's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Washington cut score on a test of equivalent difficulty. The following two figures show the difficulty of Washington's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2006 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Washington ranged between the 23rd and 49th percentiles for the norm group, with seventh grade being most challenging. In **mathematics**, the proficiency cut scores ranged between the 36th and 59th percentiles with seventh grade again being most challenging.

With the exception of fourth grade reading, Washington's cut scores in reading and mathematics are consistently at or above the median difficulty among the states studied. Note, though, that Washington's cut scores for reading are generally lower than its math cut scores. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Washington students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

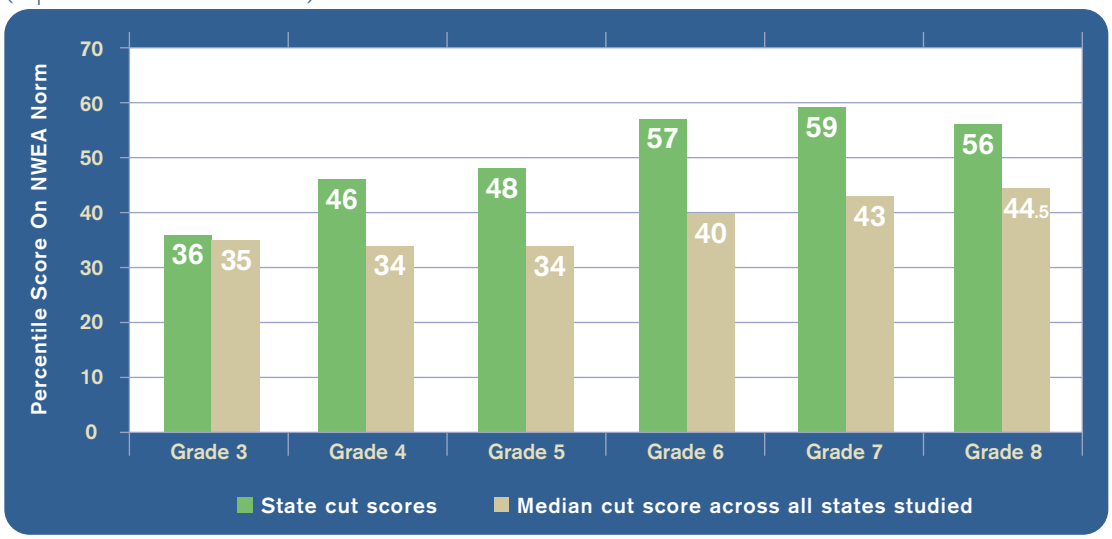
Another way of assessing difficulty is to evaluate how Washington's proficiency cut scores rank relative to other states. Table 1 shows that, except for third- and fourth-grade reading, the Washington cut scores generally rank in the middle to upper third in difficulty among the 26 states studied for this report. Its reading cut scores in grade 7 and its math cut scores in grades 7 and 8 are particularly high.

Figure 1 – Estimate of Washington Reading Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: This figure compares reading cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Washington's cut scores surpass the median cut scores in grades 3, 6, and 7, but not in the other grades.

Figure 2 – Estimate of Washington Mathematics Cut Scores in Relation to All 26 States Studied, 2006 (Expressed in MAP Percentiles)



Note: Washington's math test cut scores are shown as percentiles of the NWEA norm and compared with the median cut scores of other states reviewed in this study. Washington's cut scores surpass the median in grades 3 through 8.

Table 1 – Washington Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2006

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	5	20	17	9	3	9
Mathematics	12	5	7	5	4	4

Note: This table ranks Washington's cut scores relative to the cut scores of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Changes in Cut Scores over Time

In order to measure their consistency, Washington's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2004 and 2006 school years. Proficiency cut scores for mathematics and reading were available for both years for grades 4 and 7.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. Plus, unintentional drift can occur even in states, such as Washington, that maintained their proficiency levels.

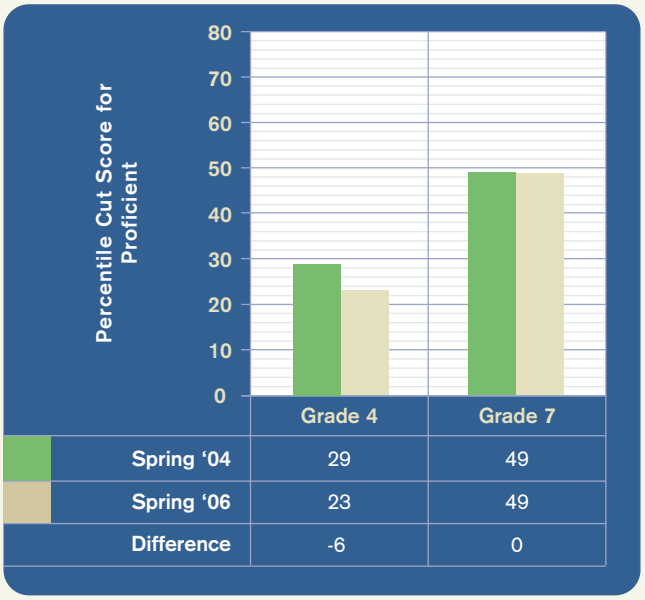
Is it possible, then, to compare the proficiency scores between the Washington's tests at these two points in time? Yes. Assume that we're judging a group of fourth graders on their high-jump ability and that we measure this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height to judge proficiency. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet, because we know the relationship between the measures. The same principle applies here. The measures or scales used by the WASL in 2004 and in 2006 can both be linked to the scale that was used to report MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can estimate the cut score needed to pass the WASL in 2004 and 2006 on the MAP scale and ascertain whether the test may have changed in difficulty. This allows us to reasonably estimate whether the WASL in 2006 is easier to pass, more difficult, or about the same as it was in 2004.

Washington's estimated **reading** cut scores indicate a decrease in difficulty over this two-year period in the fourth grade (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the fourth-grade reading proficiency rate in 2006 to be 6 percent higher than in 2004. At grade 7, there was no change in the reading proficiency cut score. (Washington reported a 7-point gain for fourth graders over this period.)

Washington's estimated **mathematics** cut scores show no substantive changes in the proficiency cut scores at fourth or seventh grades (see Figure 4). In other words, the difference in cut scores between 2004 and 2006 was less than the standard error of measurement, or 3 RIT points.

Thus, one could fairly say that Washington's fourth-grade reading test was easier to pass in 2006 than in 2004. As a result, improvements in the state's fourth-grade reading proficiency rate during this period may not be entirely a product of improved achievement. Because there were no substantive changes in the proficiency cut scores for fourth-grade math, or in either test in seventh grade, one could reasonably attribute any observed changes in proficiency ratings in these areas to actual changes in student performance.

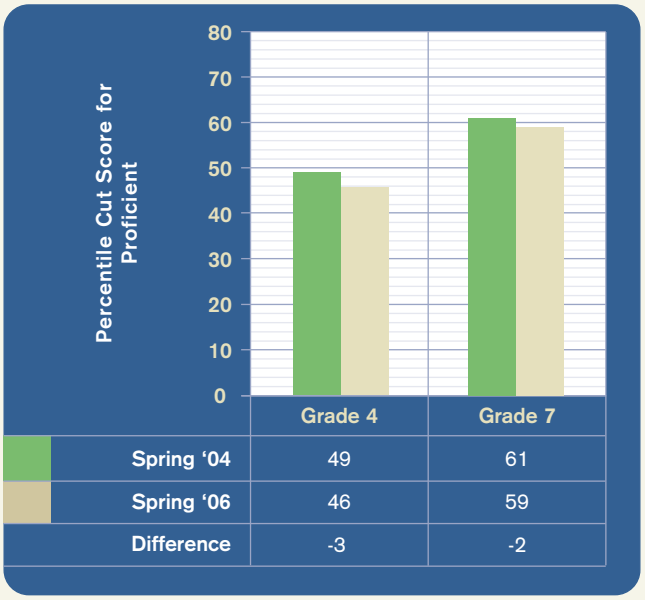
Figure 3 – Estimated Change in Washington's Proficiency Cut Scores in Reading, 2004-2006 (Expressed in MAP Percentiles)



	Grade 4	Grade 7
Spring '04	29	49
Spring '06	23	49
Difference	-6	0

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, fourth grade students in 2004 had to score at the 29th percentile of the NWEA norm group nationally in order to be considered proficient, while by 2006 fourth graders had only to score at the 23rd percentile of the NWEA norm group to achieve proficiency. The changes in grade 7 were within the margin of error (in other words, too small to be considered substantive.

Figure 4 – Estimated Differences in Washington's Proficiency Cut Scores in Mathematics, 2004-2006 (Expressed in MAP Percentiles)



	Grade 4	Grade 7
Spring '04	49	61
Spring '06	46	59
Difference	-3	-2

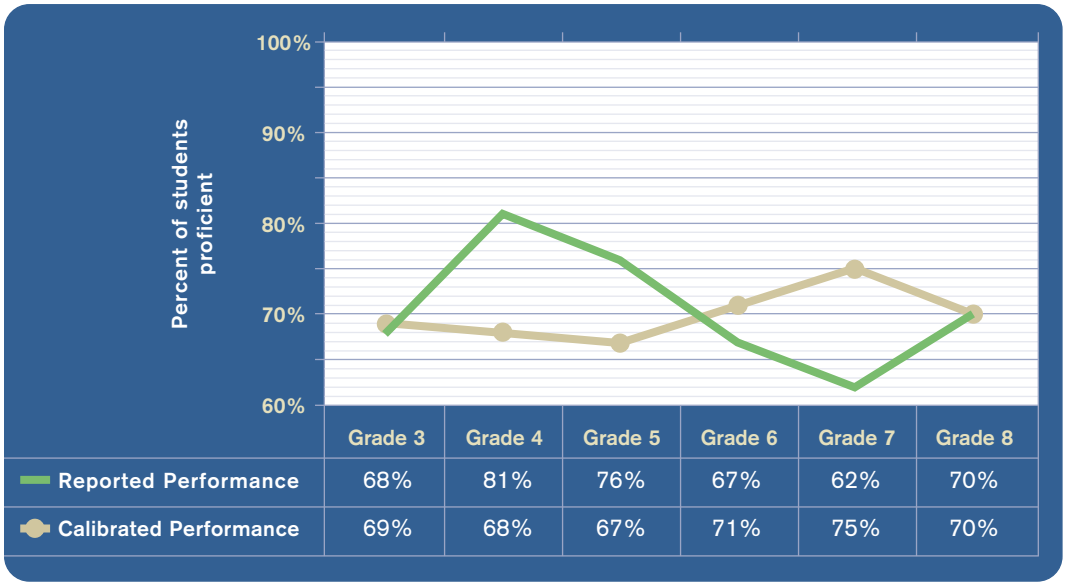
Note: This graphic shows that the difficulty of achieving proficient in math did not changed significantly. For example, fourth-grade students in 2004 had to score at the 49th percentile of the NWEA norm group nationally in order to be considered proficient, while in 2006, fourth graders had to score at the 46th percentile of the NWEA norm group to achieve proficiency—essentially no difference. The changes in both grades 4 and 7 were within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Washington's cut scores, we find that they are not well calibrated across grades. Figures 1 and 2 showed that Washington's upper-grade cut scores in reading and mathematics tended to be more challenging than the cut scores in the lower grades, particularly for mathematics. The two figures that follow show Washington's reported performance on the state test in reading (Figure 5) and mathematics (Figure 6) compared with the rate of proficiency that would be achieved if the cut scores were all calibrated to the grade-8 standard. When differences in grade-to-grade difficulty of the cut scores are removed, student performance is more consistent at all grades, especially in mathematics. This would lead to the conclusion that the higher rates of math proficiency that the state has reported for elementary school students are somewhat misleading.

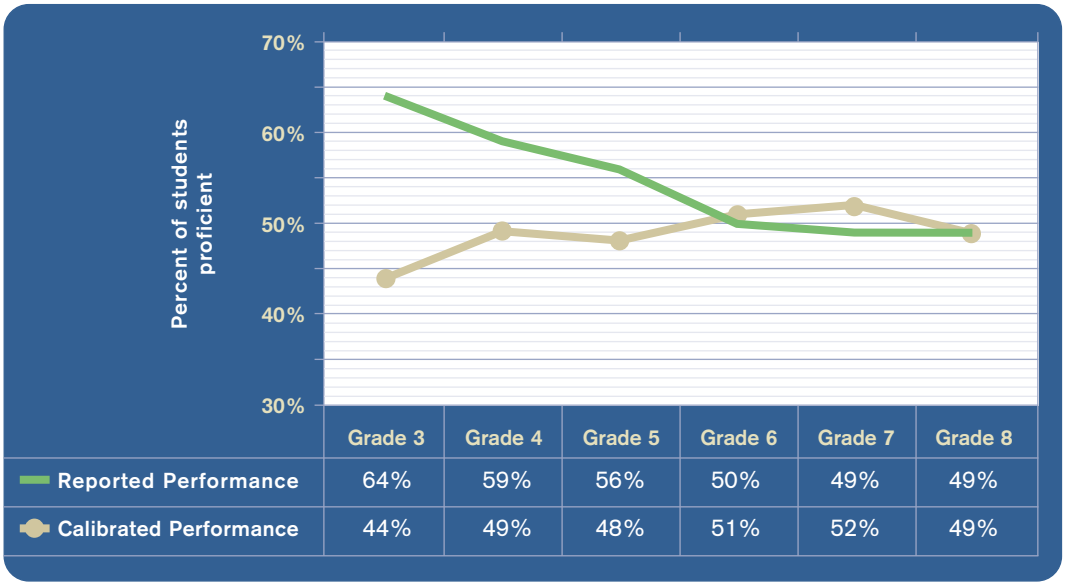
Figure 5 – Washington Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	68%	81%	76%	67%	62%	70%
Calibrated Performance	69%	68%	67%	71%	75%	70%

Note: This graphic shows, for example, that if Washington's grade-4 reading cut score was set at the same level of difficulty as its grade-8 cut score, 68 percent of fourth graders would achieve the proficient level, rather than 81 percent, as was reported by the state.

Figure 6 – Washington Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2006



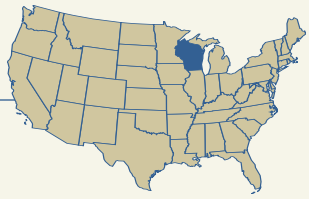
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	64%	59%	56%	50%	49%	49%
Calibrated Performance	44%	49%	48%	51%	52%	49%

Note: This graphic shows, for example, that if Washington's grade-3 mathematics cut score was set at the same level of difficulty as its grade-8 cut score, 44 percent of third graders would achieve the proficient level, rather than 64 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what constitutes proficiency in reading and math, Washington is relatively high, at least compared to the other 25 states in this study, except in grade-4 reading. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which found Washington's math standards to be in the top third and its grade-4 and grade-8 reading standards toward the middle of states studied. However, Washington's expectations are not smoothly calibrated across grades, particularly for mathematics. Students who are proficient in third grade are not necessarily on track to be proficient by eighth grade. State policymakers might consider adjusting their proficiency cut scores across grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

Wisconsin



Introduction

This study linked data from the 2003 and 2005 administrations of Wisconsin's reading and math tests to the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment, a computerized adaptive test used in schools nationwide. We found that Wisconsin's definitions of proficiency in reading and mathematics are relatively less difficult than the cut scores set by other states. In other words, Wisconsin's tests are below average in terms of difficulty.

The level of difficulty of these cut scores decreased in some grades from 2003 to 2005—the No Child Left Behind era. For example, Wisconsin's eighth-grade tests for reading and mathematics were easier in 2005 than in 2003.

Wisconsin's cut scores in mathematics are now more difficult in the lower grades than in the higher grades (taking into account the obvious differences in subject content and children's development). Consequently, the proportion of younger students who are on track to meet the cut scores at the later grades may be underestimated. Wisconsin policymakers might consider adjusting their cut scores to ensure equivalent difficulty at all grades so that parents and schools can be assured that elementary school students scoring at the proficient level are truly prepared for success later in their educational careers.

What We Studied: Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT)

Wisconsin currently uses a fall assessment called the Wisconsin Knowledge and Concepts Examinations - Criterion Referenced Test (WKCE-CRT), which tests reading, language applications, mathematics, science, and social studies in students in grades 3 through 8 and 10, as expected by NCLB. Fall 2005 was the first year the criterion-referenced test was used. It replaced the Wisconsin Knowledge and Concepts Examinations (WKCE), an augmented version of the nationally-normed Terra Nova test, first used in fall 2002 to test reading, language arts, mathematics, science, and social studies in grades 4, 8, and 10. The current study linked reading and math data from fall 2003 WKCE administrations and fall 2005 WKCE-CRT administrations to a common scale also administered in the 2003-4 and 2005-6 school years.

To determine the difficulty of Wisconsin's proficiency cut scores, we linked data from state tests to the NWEA assessment. (A "proficiency cut score" is the score a student must achieve in order to be considered proficient.) This was done by analyzing a group of elementary and middle schools in which almost all students took both the state's assessment and the NWEA test. (The methodology section of this report explains how performance on these two tests was compared.)

Part 1: How Difficult are Wisconsin's Definitions of Proficiency in Reading and Math?

One way to evaluate the difficulty of a standard is to determine how many people attempting to attain it are likely to succeed. How do we know that a two-foot high bar is easy to jump over? We know because if we asked 100 people at random to attempt such a jump, perhaps 80 would make it. How do we know that a six-foot high bar is challenging? Because only one (or perhaps none) of those same 100 individuals would successfully meet that challenge. The same principle can be applied to academic standards. Common sense tells us that it is more difficult for students to solve algebraic equations with two unknown variables than it is for them to solve an equation with only one unknown variable. But we can figure out exactly how much more difficult by seeing how many eighth graders nationwide answer both types of questions correctly.

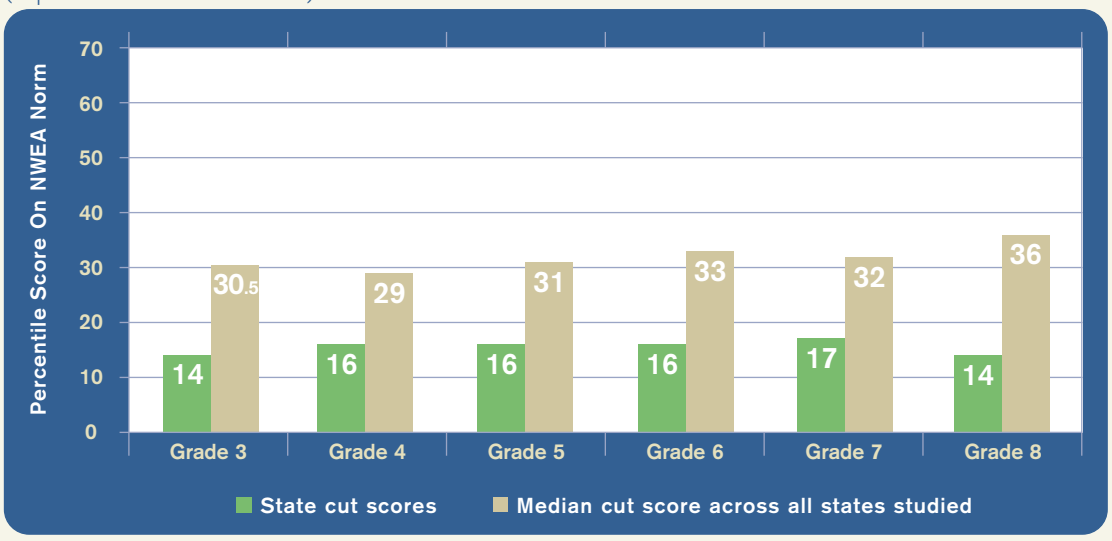
Applying that approach to this assignment, we evaluated the difficulty of Wisconsin's proficiency cut scores by estimating the proportion of students in NWEA's norm group who would perform above the Wisconsin cut score on a test of equivalent difficulty. The following two figures show the difficulty of Wisconsin's proficiency cut scores for reading (Figure 1) and mathematics (Figure 2) in 2005 in relation to the median cut score for all the states in the study. The proficiency cut scores for **reading** in Wisconsin ranged between the 14th and 17th percentiles for the norm group, with the seventh-grade cut score being most challenging. In **mathematics**, the proficiency cut scores ranged between the 21st and 29th percentiles with the third and fourth grade cut scores being most challenging.

For all grade levels, Wisconsin's cut scores in both reading and mathematics are lower than the median cut scores of the other states in the study, and far below the capabilities of the average student of that grade within the NWEA norm group.

Note, too, that Wisconsin's cut scores for reading are lower than those for mathematics. Thus, reported differences in achievement between the two subjects may be more a product of differences in cut scores than in actual student achievement. In other words, Wisconsin students may be performing worse in reading and better in mathematics than is apparent by just looking at the percentage of students passing state tests in those subjects.

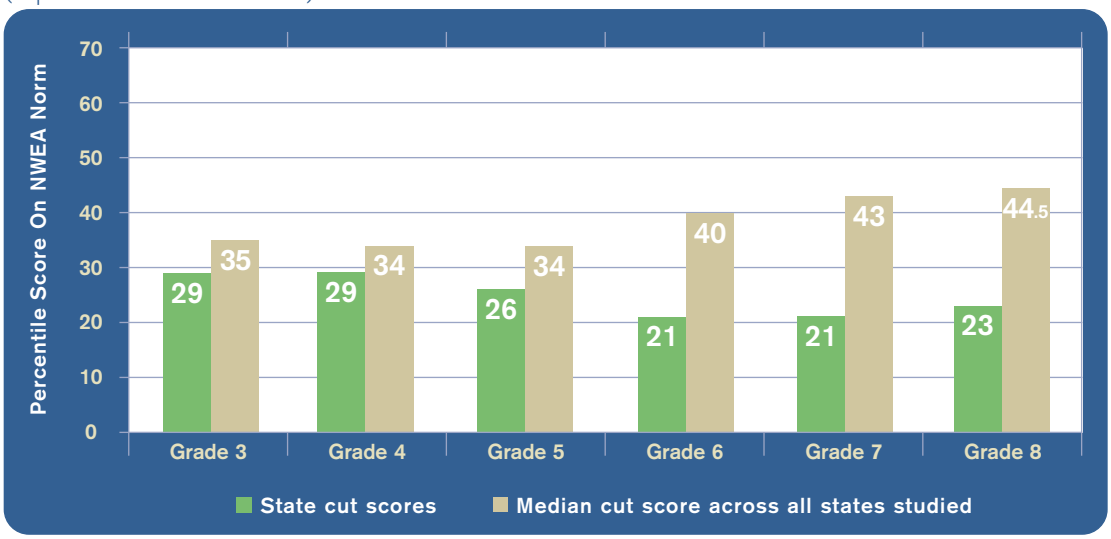
Another way of assessing difficulty is to observe how Wisconsin's proficiency cut scores rank relative to other states. Table 1 shows that the state's cut scores generally rank among the lowest of the 26 states studied for this report, in terms of difficulty.

Figure 1 – Wisconsin Reading Cut Scores in Relation to All 26 States Studied, 2005 (Expressed in MAP Percentiles).



Note: This figure compares reading test cut scores ("proficiency passing scores") as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Wisconsin's scores range from 13 to 22 percentile points behind the median.

Figure 2 – Wisconsin Mathematics Cut Scores in Relation to All 26 States Studied, 2005
(Expressed in MAP Percentiles)



Note: This figure compares reading test cut scores as percentiles of the NWEA norm. These percentiles are compared with the median cut score of all 26 states reviewed in this study. Wisconsin's scores range from 5 to 22 percentile points behind the median.

Table 1 – Wisconsin Rank for Proficiency Cut Scores Among 26 States in Reading and Mathematics, 2005

	Ranking (Out of 26 States)					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading	23	24	23	24	25	23
Mathematics	19	18	18	23	23	21

Note: This table ranks Wisconsin's cut scores relative to those of the other 25 states in the study, with 1 being highest and 26 lowest.

Part 2: Changes in Cut Scores over Time

In order to measure their consistency, Wisconsin's proficiency cut scores were mapped to their equivalent scores on NWEA's MAP assessment for the 2003-4 and 2005-6 school years during the same season. Cut score estimates for reading and mathematics were available for both years in grades 4 and 8.

States may periodically re-adjust the cut scores they use to define proficiency in reading and math, or may update the tests used to measure student proficiency. Such changes can impact proficiency ratings, not necessarily because student performance has changed, but because the measurements and criteria for success have changed. This was the case for Wisconsin which, as explained above, adopted a new test for 2005.

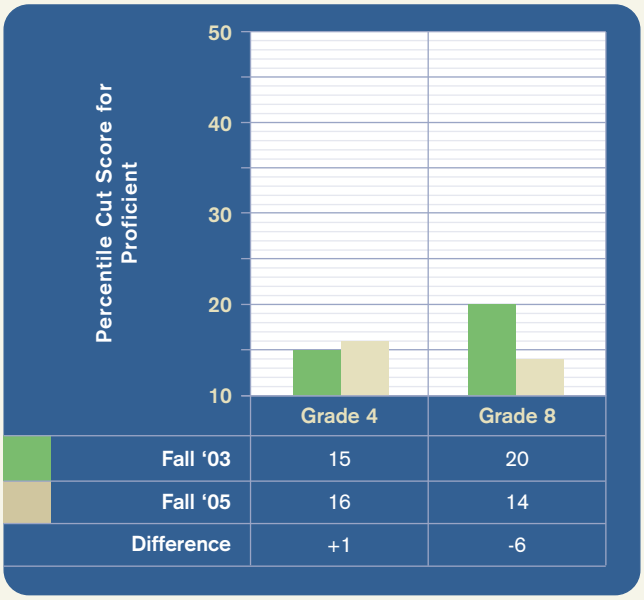
Is it possible, then, to compare the proficiency scores between the earlier and later administrations of Wisconsin tests? Yes. Assume that we're judging a group of fifth graders on their high-jump prowess and that we gauge this by finding how many in that group can successfully clear a three-foot bar. Now assume that we change the measure and set a new height. Perhaps students must now clear a bar set at one meter. This is somewhat akin to adjusting or changing a state test and its proficiency requirements. Despite this, it is still possible to determine whether it is more difficult to clear one meter than three feet because we know the relationship between the measures. The same principle applies here. The measures or scales used by the WKCE in 2003 and the WKCE-CRT in 2005 can both be linked to the MAP, which has remained consistent over time. Just as one can compare three feet to one meter and know that a one-meter jump is slightly more difficult than a three-foot jump, one can use the MAP scale to estimate whether the WKCE-CRT in 2005 is easier or more difficult than the prior test and proficiency cut scores that were in place.

In **reading**, Wisconsin showed a moderate decrease in the estimated eighth-grade reading cut score estimate over this two-year period, but essentially no change in the fourth-grade reading cut score (see Figure 3). Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the eighth-grade reading proficiency rate in 2005 to be 6 percent higher than in 2003. (In fact, Wisconsin reported a 6-point gain for eighth graders over this period.)

Wisconsin's **mathematics** results show the same pattern, with a moderate decrease in the estimated eighth-grade cut score and essentially no change in the fourth-grade cut score. Consequently, even if student performance stayed the same on an equivalent test like NWEA's MAP assessment, one would expect the eighth-grade math proficiency rate in 2005 to be about 11 percent higher than in 2003, even if actual student performance remained the same. (Wisconsin Wisconsin reported a 9-point gain for eighth graders over this period.)

Thus, one could fairly say that Wisconsin's fourth-grade tests in both reading and mathematics stayed about the same from 2003 to 2005, while the eighth-grade tests became easier to pass. As a result, improvements in state-reported proficiency rates during this period may not be entirely a product of improved achievement.

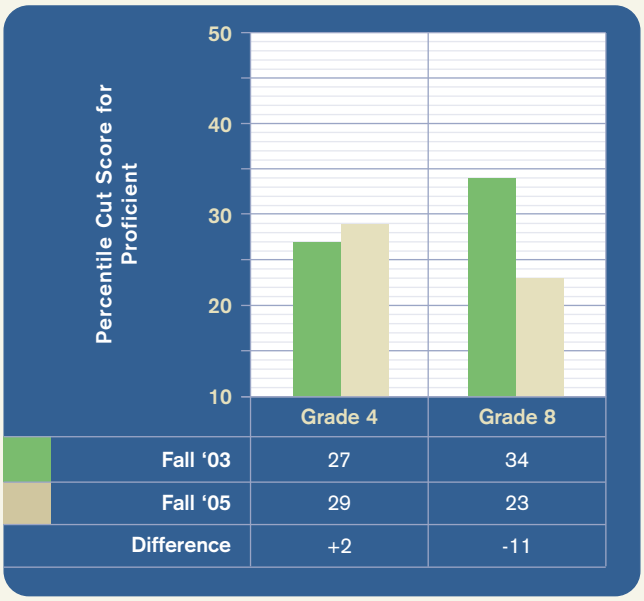
Figure 3 – Estimated Differences in Wisconsin's Proficiency Cut Scores in Reading, 2003-2005 (Expressed in MAP Percentile Ranks)



	Grade 4	Grade 8
Fall '03	15	20
Fall '05	16	14
Difference	+1	-6

Note: This graphic shows how the difficulty of achieving proficiency in reading has changed. For example, eighth-grade students in 2003 had to score at the 20th percentile nationally in order to be considered proficient, while by 2005 eighth graders had to score at the 14th percentile to achieve proficiency. The change in grade 4 was within the margin of error (in other words, too small to be considered substantive)

Figure 4 – Estimated Differences in Wisconsin's Proficiency Cut Scores in Mathematics, 2003-2005 (Expressed in MAP Percentiles)



	Grade 4	Grade 8
Fall '03	27	34
Fall '05	29	23
Difference	+2	-11

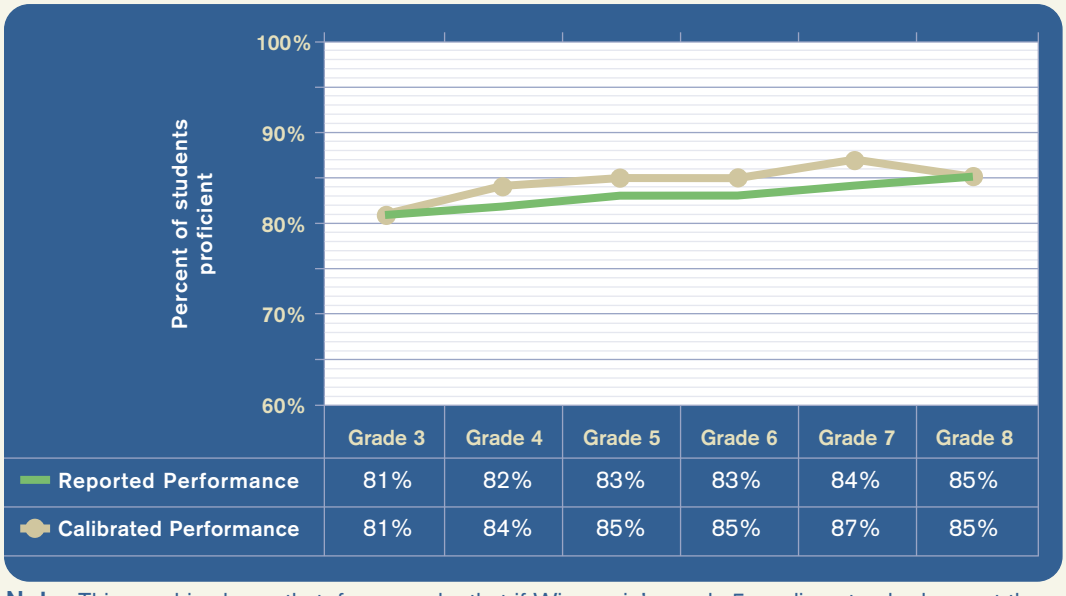
Note: This graphic shows how the difficulty of achieving proficiency in math has changed. For example, eighth-grade students in 2003 had to score at the 34th percentile nationally in order to be considered proficient, while in 2005 eighth graders only had to score at the 23rd percentile of the NWEA norm group to achieve proficiency. The change in grade 4 was within the margin of error (in other words, too small to be considered substantive).

Part 3: Calibration across Grades

Calibrated proficiency cut scores are those that are relatively equal in difficulty across all grades. Thus, an eighth-grade cut score would be no more or less difficult for eighth graders to achieve than a third-grade cut score is for third graders. When cut scores are so calibrated, parents and educators have some assurance that achieving the third-grade proficiency cut score puts a student on track to achieve the standards at eighth grade. It also provides assurance to the public that reported differences in performance across grades are a product of differences in actual educational attainment and not simply differences in the difficulty of the test.

Examining Wisconsin's cut scores, we see in Figures 1 and 2 showed that the state's reading cut scores across grades 2 through 8 were fairly well calibrated, while the math cut scores in the lower grades were slightly more difficult than in the upper grades. These are reflected in Figures 5 and 6, which show how Wisconsin's reported performance on the state test in reading (Figure 5) and mathematics (Figure 6) compared with the rate of proficiency that would be achieved if the cut scores were all calibrated to the eighth-grade standard. In Figure 5, the differences between the observed proficiency rates and those that would be expected with calibrated cut scores are quite small. In Figure 6, however, we see that the uncalibrated standards at the earlier grades slightly underestimate the proportions of third and fourth graders who are on track to eventually demonstrate proficiency at the later grades.

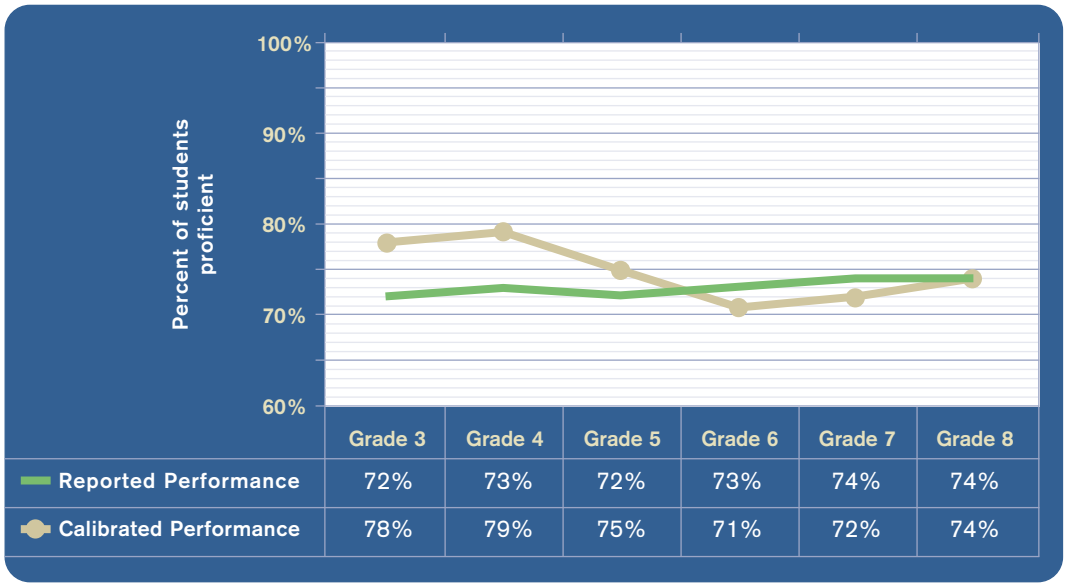
Figure 5 – Wisconsin Reading Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	81%	82%	83%	83%	84%	85%
Calibrated Performance	81%	84%	85%	85%	87%	85%

Note: This graphic shows that, for example, that if Wisconsin's grade-5 reading standard was at the same difficulty level as its grade-8 standard, 85 percent of fifth graders would achieve the proficient level, rather than 83 percent, as was reported by the state.

Figure 6 – Wisconsin Mathematics Performance as Reported and as Calibrated to the Grade-8 Standard, 2005



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reported Performance	72%	73%	72%	73%	74%	74%
Calibrated Performance	78%	79%	75%	71%	72%	74%

Note: This graphic shows, for example, that if Wisconsin's grade-3 mathematics cut score was set at the same difficulty level as its grade-8 cut score, 78 percent of third graders would achieve the proficient level, rather than 72 percent, as was reported by the state.

Policy Implications

When setting its cut scores for what students must know and be able to do to be considered proficient in reading and math, Wisconsin is low, compared with the other 25 states in this study. This finding is consistent with the recent National Center for Education Statistics report, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, which also found Wisconsin to have some of the lowest standards of all states, at least in reading. In the past several years, the difficulty of the grade-8 cut scores has declined somewhat. As a result, Wisconsin's expectations for mathematics are not smoothly calibrated across grades, so Wisconsin currently underestimates the proportion of students in the younger grades who are on track to meet the (low) eighth-grade mathematics cut scores. Wisconsin policymakers might consider adjusting their cut scores across grades so that proficiency at the earlier grades more accurately predicts proficiency at the later grades.

Appendix 1 - Methodology

This study used data collected from schools whose students participated in both state testing and in the Measures of Academic Progress (MAP) assessment of the Northwest Evaluation Association (NWEA) (Northwest Evaluation Association 2003). Its purpose was to estimate the proficiency cut scores for twenty-six state assessments, using the NWEA scale as a common ruler. For nineteen of those states, estimates of cut scores could be made at two points in time, and these were used to monitor any changes that occurred during the process of implementing the No Child Left Behind Act (NCLB) requirements.

Instruments

Proficiency results from state assessments offered in grades 3 through 8 in reading or English/language arts and in mathematics were linked to reading and mathematics results on NWEA's MAP tests. MAP tests are computer-adaptive assessments in the basic skills covering grade 2 through high school that are taken by students in about 2,570 school systems in forty-nine states.

MAP assessments have been 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.

Validity

The notion of test validity generally refers to the degree to which a test or scale actually measures the attribute or characteristic we believe it to measure. In this case, the traits measured are mathematics achievement and reading or English/language arts achievement. The various state assessments and MAP are both instruments designed to provide a measurement of these domains. Of course, neither MAP nor the various state assessments definitively measure the underlying trait, and for purposes of this study we can only offer evidence of MAP's appropriateness for this task.

Content Validity

Content validity refers to "the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured" (Anatasi and Urbina 1997). A test has content validity built into it by careful selection of which items to include (Anatasi and Urbina 1997).

Each MAP assessment is developed from a large pool of items in each subject that have been calibrated for their difficulty to an equal-interval, cross-grade scale called the RIT scale. These pools contain approximately fifty-two hundred items in reading and eight thousand items in mathematics. Each item is aligned to a subject classification index for the content being measured. From this large pool of items, NWEA curriculum experts create a state-aligned test by reviewing the state standards and matching that structure to a highly specific subject classification index used to organize the content of the MAP item pool. From this match a subset of about two thousand items corresponding to the content standards of each state is selected. The processes governing item writing and test creation are more specifically outlined in NWEA's *Content Alignment Guidelines* (2007).

Business organizations often characterize processes like the one used to create MAP assessments as "mass customization," because they employ a single set of procedures to create products with differing individual specifications—in this case multiple tests, each of which is unique to the state in which it is used. Because the items used to create each unique state assessment come from the same parent—that is, a single item pool with all questions evaluated on a common scale—the results of various state MAP assessments can be compared to one another. MAP's alignment to each state's content standards distinguishes it from National Assessment of Educational Progress (NAEP) and other national standardized tests, such as the Iowa Test of Basic Skills, that are not aligned to state standards but instead reflect the same content across all settings in which they are used.

Each student taking MAP receives a unique test of forty to fifty-five items containing a balanced sample of items testing

the four to eight primary standards in his or her state's curriculum. The assessment is adaptive in design, so that the items given to students will closely reflect their current performance rather than their current grade. More importantly, because each test differs, MAP assessments will generally provide a broader, more diverse sampling of the state's standards than can be achieved when a single version of an assessment is offered to all students in a state.

For purposes of NCLB, the states have the discretion to test reading as a stand-alone subject or to integrate the assessment of reading into a broader test that also measures writing and language usage skills. NWEA offers separate assessments in reading and language usage and does not typically offer assessments in writing. In states that assessed the broader English/language arts domain, NWEA aligned the state test with the MAP reading assessment score, and did not attempt to combine reading and language usage scores. This practice reduced the content alignment in some cases. However, prior studies found that it did not degrade the ability of the MAP test to produce a cut score that would effectively predict proficiency on state tests using a language arts test, compared to states using a reading-only assessment (Cronin, Kingsbury, Dahlin, and Bowe 2007; NWEA 2005b). Of the twenty-six states studied here, NWEA reading tests were linked to an English/language arts assessment in four: California, Indiana, New Jersey, and South Carolina. The remaining twenty-two states all tested reading.

Concurrent Validity

Concurrent validity studies are generally employed to establish the appropriateness of using one assessment to project cut score equivalencies onto another instrument's scale. Concurrent validity is critical when trying to make predictions from one test about a student's future performance on another test. NWEA has previously published results from concurrent validity studies using MAP and fourteen state assessments that were conducted between 2002 and 2006 (Cronin et al. 2007; NWEA 2005b). These generally show strong predictive relationships between MAP and the state assessments (see Appendix 2). Across the reading studies, Pearson correlations between MAP and the fourteen state assessments averaged .79; the average correlation across the mathematics studies was .83. This is sufficient concurrent validity to suggest that results on MAP will predict results on the state assessment reasonably well.

Measurement Scale

NWEA calibrates its tests and items using the one-parameter logistic IRT model known as the Rasch model (Wright 1977). Results are reported using a cross-grade vertical scale called the RIT scale to measure student performance and growth over time. The original procedures used to derive the scale are described by Ingebo (1997). These past and current scaling procedures have two features designed to ensure the validity and stability of the scale:

1. The entire MAP item pool is calibrated according to the RIT scale. This ensures that all state-aligned tests created from the pool measure and report on the same scale. There is no need to equate forms of tests, because each derived assessment is simply a subset of a single pre-calibrated pool.

2. Ingebo employed an interlocking field test design for the original paper version of MAP, ensuring that each item was calibrated against items from at least eight other field test forms. This interlocking design resulted in a very robust item pool with calibrations that have remained largely constant for over twenty years, even as these items have transferred from use on paper-and-pencil assessments to computer-delivered assessments (Kingsbury 2003).

These procedures permit the creation of a single scale that accurately compares student performance across separate state curriculum standards. Because of the stability of the scale over time, formal changes in the state-test cut score will generally be reflected by changes in the estimated equivalent score on the RIT scale. The RIT scale estimates may also change when factors exist that change performance on a state assessment without comparably changing the NWEA assessment. For example, if a state test were changed from low stakes for students to high stakes, it is possible that student performance on the state test would improve because of higher motivation on the part of students, but MAP results would probably not change. This would cause the MAP estimated cut score for the state test to decline because students with lower scores would more frequently score proficiently on the state test. Other factors that can influence these estimates include increased student familiarity with the format and content of a test, as well as issues in the equating of state-test measurements scales that may cause drift in a state test's difficulty over time.

Sample

We computed proficiency cut score estimates for twenty-six state assessments. (The states involved are home to school districts that use the NWEA assessment.) In order to create the population samples within each state that were used to estimate these cut scores, one of two procedures was applied. Each of the two procedures produced populations of students who had taken both their respective state assessment and MAP.

When NWEA had direct access to individual student results on both the state assessment and MAP, a sample was created by linking each student's state test results to his or her RIT score using a common identification number (method 1). This resulted in a sample containing only students who had taken both tests. Proficiency cut scores for eleven states were estimated using this method.

We used the alternate procedure (method 2) when NWEA did not have individual student results from the state assessment available. This procedure matched school-level results on the state test with school-level performance on NWEA's test to estimate scores. To do this we extracted results from schools in which the count of students taking MAP was, in the majority of cases, within 5 percent of the count taking the respective state test. When matching using this criterion did not produce a sufficiently large sample, we permitted a match to within 10 percent of the count taking the respective state test.

Below are the specific steps involved in method 2:

- All valid student test records for Northwest Evaluation Association clients in the target state for the appropriate term were extracted, and their results were aggregated by school, grade, and test measurement scale.

- Data were captured from department of education websites in each state showing the number of students tested in each school and the proportion of students tested who performed at each proficiency level on the state test.

- National Center for Educational Statistics (NCES) school identification information was used to link results from the state test reports to the appropriate school reports in the NWEA database.

- The linked data sets were filtered to find schools in which the number of students who had taken the NWEA assessment was within 5 percent of the number taking the respective state exams. If this method generated at least seven hundred students per grade (the minimum we would accept) for each test measurement scale, we did not expand the study group further. If the initial criterion failed to generate that number, we liberalized the criterion to 7.5 percent[3] and finally to 10 percent. If the liberalized criterion did not identify seven hundred matches, then that grade level was removed from the study. Appendix 3 identifies the states included in the final study for mathematics and the criterion applied to achieve the necessary number of matching records.

Method 2 resulted in the identification of a group of schools in fifteen states in which nearly all students had taken both their state assessment and MAP. Because the two tests are highly correlated and reasonably aligned (see Appendix 2), this procedure produced sufficiently large matched samples to provide proficiency cut score estimates on the MAP scale that fairly represent the level of performance required to achieve proficiency on the state assessments.

During the period studied, NWEA was the provider for Idaho's state assessment, which is reported on the RIT scale. Results for Idaho, therefore, represent the actual RIT values of the past and current cut scores rather than estimates. Cut score estimates for the New England Common Assessment Program, which is used as the NCLB assessment in the states of New Hampshire, Rhode Island, and Vermont, were derived from a sample of New Hampshire students.

These procedures produced proficiency cut score estimates for twenty-six states. Of these, nineteen produced cut scores for multiple test years, allowing us to examine changes over time.

[3] An analysis was conducted to determine whether the more liberal 10 percent inclusion criterion could introduce any bias into the estimated cut scores. A small biasing effect was found, resulting in estimated cut scores that were, on average, 0.3 raw scale units higher than were generated using the more stringent inclusion criterion. In no single case was the difference in the cut score estimate larger than the standard error of measurement. The small bias introduced by the 10 percent inclusion criterion had no discernable effects on the corresponding percentile scores for a given cut score estimate.

Estimates Part 1:
Proficiency Cut Scores in Reading and Math

The sampling procedures identified populations in which nearly all students took both their respective state assessment and the NWEA assessment. To estimate proficiency level cut scores, we calculated the proportion of students in the sample population who performed at a proficient or above level on the state test and then found the minimum score on the RIT scale from the rank-ordered MAP results of the sample that would produce an equivalent proportion of students. This is commonly referred to as an equipercentile method of estimation. Thus, if 75 percent of the students in the sample achieved proficient performance on their state assessment, then the RIT score of the 25th percentile student in the sample (100 percent of the group minus the 75 percent of the group who achieved proficiency) would represent the minimum score on MAP associated with proficiency on the state test.

This equipercentile or "distributional" method of estimation was chosen pursuant to a study of five states conducted by Cronin and others (2007). This study compared the accuracy of proficiency level estimates derived using the equipercentile methodology to estimates that were derived from prior methods used by NWEA to link state assessment cut scores to the RIT scale. These prior methods included three techniques to estimate cut scores: linear regression, second-order regression, and Rasch status-on-standard modeling. The study found that cut score estimates derived from the equipercentile methodology came the closest to predicting the actual state assessment results for the students studied. In mathematics, compiled MAP proficiency estimates overpredicted the percentage of students who were proficient on state tests by only 2.2 percentage points on average. In the reading domain, compiled MAP proficiency estimates overpredicted actual state test results by about 3 percent on average across the five states. This level of accuracy was deemed sufficient to permit reasonable estimates of the difficulty of state assessments and general comparisons of the difficulty of proficiency cut scores across states in the two domains studied.

Once the proficiency cut scores were estimated on the RIT scale, they were converted to percentile scores in order to permit comparisons across states that tested students during different seasons. When possible, averages or other summary statistics reported as percentile scores in this study were first calculated as averages of scale scores, and then converted to their percentile rank equivalent. The MAP percentile scores reported come from NWEA's most recent norming study (NWEA 2005b). The norming sample was composed of over 2.3 million students who attended 5,616 schools representing 794 school systems in 32 states. All school systems that had tested with NWEA for longer than one year were invited to participate in the study. NWEA included all valid, official test results for those school systems for the fall and spring terms of 2003 and 2004. Because all volunteering school systems were included, the sample was selected to represent as broad a cross-section of the large NWEA testing population as possible, and was not intended to reflect the geographic and ethnic distribution of the United States as a whole. In an effort to determine whether the performance of the normative sample differed from a sample representing the nation's ethnic balance, results from the normative sample were later compared to a smaller sample from the NWEA testing population that was selected for balance on this trait. These analyses were reported as part of the norms study. Mean scale score differences between these two samples were less than 1.5 scale score points across all grades and subjects (Northwest Evaluation Association 2005b). These differences were small enough to suggest that the norm group sample produced results that did not differ significantly from a sample representative of the ethnic makeup of the population of school-age children in the United States.

Estimates Part 2:
Changes in Cut Scores over Time

Multiple estimates were generated for twenty states, permitting comparisons of cut scores over time. The most recent estimate was taken from data gathered during the spring 2005, fall 2005, spring 2006, fall 2006, or spring 2007 testing term. The initial estimate was taken from the oldest term between spring 2002 and spring 2005 that would produce an adequate sample.

Estimates Part 3:
Calibration across Grades

One purpose of academic standards is to set expectations for performance that are transparent and consistent across a course of study. For standards to be consistent, we believe, the difficulty of the standard should be similar or calibrated across all grades in school.

Assume, for example, that a third-grade reading proficiency standard was established at a level that was achieved by 70 percent of all third-graders within a large norming sample. Now assume that an eighth-grade reading standard was also established that could be met by 70 percent of all eighth-graders in the same large norming sample. We would say that these two standards are calibrated, or equivalent in terms of relative difficulty, since the same proportion of students (70 percent) in the norming samples successfully mastered both standards.

Armed with the knowledge that these third- and eighth-grade standards are calibrated, let us now assume that a state using these standards reports that 60 percent of its third-grade students achieved the third-grade standard, while 80 percent of its eighth-grade students achieved the eighth-grade standard. Because the standards are calibrated, we know that the reported differences between third- and eighth-grade achievement represent true differences in student performance and not differences in the relative difficulty of the tests.

Because NCLB requires testing of students in grades 3 through 8, eighth grade was selected as the end point for purposes of estimating calibration. By comparing the NWEA norm group percentile scores associated with the standard at each grade, we were able to determine how closely they were calibrated, relative to the difficulty level of the standard at the end of middle school.

When proficiency standards are calibrated, successful performance at one grade will predict successful performance at a later grade, assuming the student continues to progress normally. A third-grade learning standard, for example, does not exist for its own sake, but represents the level of skill or mastery a student needs if he or she is to go on to meet the challenges of fourth-grade. In other words, the standards at each grade exist to ensure that students have the skills necessary to advance to the next level.

Non-calibrated standards do not prepare students to meet future challenges, particularly when the standards at the earliest grades are substantially easier than the standards at the later grades. If a third-grade standard is sufficiently easy that third-graders can achieve it with only a modest amount of effort, then those students are not being adequately prepared to meet future standards, which might require significantly more effort.

Students with sufficient skill to meet a very easy standard might not have the ability to meet a more difficult standard. Consequently, one would expect that the percentage of students who meet their state's proficiency requirements would be higher when the standard is easy, and lower when the standard is difficult. Indeed, it is possible to quantify the degree of impact on the state proficiency ratings attributable to non-calibrated standards when expressing state standards as percentile rankings.

To illustrate this process, we will use the MAP proficiency cut score estimates for the 2005 Arizona state assessment (AIMS) in mathematics. We estimated the AIMS proficiency standard at eighth grade to be at the 42nd percentile of the NWEA norm group for this grade, meaning that 58 percent of the norm group would be likely to perform above this standard. The standard at third grade, however, is lower. It is set at the 30th percentile on NWEA norms, which means that 70 percent of the norm group would be likely to perform above this standard. To use simple math, we estimated that this difference in the difficulty of the cut scores would cause 12 percent more students to pass the third-grade standard than the eighth-grade standard (see Table A1.1). Next, we extracted the actual results reported for the 2005 AIMS assessment. These results show that 77 percent of Arizona students passed the third-grade test. As expected, a smaller number, 63 percent, passed the eighth-grade exam.

Table A1.1 – NWEA percentile scores associated with proficient performance on Arizona AIMS in mathematics - 2005

	Grade 3	Grade 8	Difference
Percentile score	30th	42nd	-12

The question is whether the difference between the third- and eighth-grade mathematics achievement is primarily a product of differences in student achievement, or a reflection of the differences in the difficulty of the test. To remove the impact of difficulty on reported achievement, we simply subtracted the differences in performance attributable to differences in the difficulty of the test (in the current example, 12 percent) from the state's reported proficiency rates on the test. The result (see Table A1.2) shows that third- and eighth-graders performed nearly the same after accounting for differences in the difficulty of the cut score.

Table A1.2 – Estimated Arizona AIMS performance in mathematics after adjusting for differences in proficiency cut score difficulty

	Grade 3	Grade 8
State-reported proficiency rating (pass rate)	77%	63%
Difference from 8th grade (from A1.1 above)	-12%	0%
Adjusted (calibrated) pass rate	65%	63%

The three parts of this appendix dealing with estimates have provided descriptions and details of the methods used to estimate proficiency cut scores within and across differing state tests and test subject areas. Each part provided the details that permitted us to answer the three major questions in the study: 1) How consistent are the various states' expectations for proficiency in reading and mathematics? 2) Is there evidence that states' expectations for proficiency have changed over time? 3) How closely are proficiency standards calibrated across grades? That is, are the standards in earlier grades equal in difficulty to proficiency standards in later grades?

Appendix 2 - Summary of Concurrent Validity Studies

Table A2.1 – Correlation between state reading or English/language arts tests and Northwest Evaluation Association's Measures of Academic Progress

Assessment	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Average
Arizona (AIMS) 2006*	0.85	0.82	0.83	0.82	0.81	0.80	0.82
California (CST) 2003*	0.84	0.83	0.83	0.82	0.83	0.83	0.83
Colorado (CSAP) 2006	0.81	0.84	0.86	0.88	0.88	0.87	0.86
Delaware (DSTP) 2006	0.76	0.76	0.75	0.74	0.78	0.78	0.76
Illinois (ISAT) 2003	0.80		0.80			0.79	0.80
Michigan (MEAP) 2006	0.76	0.78	0.77	0.77	0.75	0.77	0.77
Minnesota (MCA & BST) 2003	0.82		0.83			0.77	0.81
Montana (MontCAS) 2004		0.82				0.79	0.81
Nevada (CRT) 2003	0.82		0.83				0.83
New Hampshire (NECAP) 2006	0.82	0.79	0.74	0.79	0.79	0.71	0.77
South Carolina (PACT) 2003*	0.76	0.79	0.78	0.77	0.78	0.76	0.77
Pennsylvania (PSSA) 2003			0.84			0.84	0.84
Texas (TAKS) 2003	0.66		0.70	0.72	0.69		0.69
Washington (WASL) 2004		0.77			0.78		0.78
Count	11	9	12	8	9	11	14
Average	0.79	0.80	0.80	0.79	0.79	0.79	0.80

* Indicates reading test was correlated to an English/language arts test

Table A2.2 – Correlation between state and norm-referenced mathematics tests and Northwest Evaluation Association's Measures of Academic Progress

Assessment	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Average
Arizona (AIMS) 2006	0.84	0.85	0.86	0.87	0.87	0.88	0.86
California (CST) 2003	0.82	0.83	0.84	0.86	0.85	0.77	0.83
Colorado (CSAP) 2006	0.81	0.84	0.86	0.88	0.88	0.87	0.86
Delaware (DSTP) 2006	0.81	0.85	0.81	0.85	0.87	0.85	0.84
Illinois (ISAT) 2003	0.8		0.8			0.79	0.80
Michigan (MEAP) 2006	0.78	0.81	0.84	0.83	0.84	0.83	0.82
Minnesota (MCA & BST) 2003	0.77		0.83			0.85	0.82
Montana (MontCAS) 2004		0.75				0.84	0.80
Nevada (CRT) 2003	0.76		0.86				0.81
New Hampshire (NECAP) 2006	0.82	0.84	0.85	0.87	0.86	0.88	0.85
South Carolina (PACT) 2003	0.76	0.84	0.84	0.84	0.85	0.85	0.83
Pennsylvania (PSSA) 2003			0.87			0.85	0.86
Texas (TAKS) 2003			0.76		0.82		0.79
Washington (WASL) 2004		0.78			0.88		0.83
Count	10	9	12	7	9	11	14
Average	0.80	0.82	0.84	0.86	0.86	0.84	0.83

Appendix 3

Tables A3.1–mathematics and A3.2–reading summarize key information about each of the state alignment studies, showing the year and school term in which the study was conducted, the grades evaluated, and the average number of students in each grade included. The tables show whether the estimate was derived directly, using a group of students who had taken both MAP and their respective state assessment, or indirectly, using cumulative MAP and state test results from schools in which nearly all students were known to have taken both tests. When the indirect method was used, the match level shows how closely the count of students testing on MAP matched the count of students taking the state test. For example, 95 percent to 105 percent would mean that the count of students taking MAP was between 95 percent and 105 percent of the count of students taking the state assessment.

Table A3.1 – Summary of Study Method and Sample Population by State - **Mathematics**

State	Term	Method	Grades	Average student count per grade	Match Level
AZ	Spring 02	1	3, 5, 8	2408	--
AZ	Spring 05	1	3,4,5,6,7,8	2828	--
CA	Spring 03	1	3,4,5,6,7,8	9257	--
CA	Spring 06	1	3,4,5,6,7	8449	95% - 105%
CO	Spring 02	1	5,6,7,8	6075	--
CO	Spring 05	1	3,4,5,6,7,8	3115	--
DE	Spring 06	2	3,4,5,6,7,8	2107	--
ID	Spring 03	NWEA administered state test	3,4,5,6,7,8	--	--
ID	Spring 06	NWEA administered state test	3,4,5,6,7,8	--	--
IL	Spring 03	1	3,5,8	1654	--
IL	Spring 06	1	3,4,5,6,7,8	1179	--
IN	Fall 02	1	3,6,8	2695	--
IN	Fall 06	2	3,4,5,6,7,8	13796	95% - 105%
KS	Fall 06	1	3,4,5,6,7,8	2365	--
MA	Spring 06	2	3,4,5,6,7,8,10	1605	92.5% - 107.5%
ME	Spring 06	2	3,4,5,6,7,8	1597	95% - 105%
MI	Fall 03	2	4,8	1637	92.5% - 107.5%
MI	Fall 05	1	3,4,5,6,7,8	2479	--
MN	Spring 03	1	3,5,8	4363	--
MN	Spring 06	1	3,4,5,6,7,8	19718	--
MT	Spring 04	1	4,8,10	1412	--
MT	Spring 06	2	3,4,5,6,7,8	1984	95% - 105%
ND	Fall 04	1	3,4,5,6,7,8	1527	--
ND	Fall 06	2	3,4,5,6,7,8	1890	90% - 110%
NH	Fall 03	2	3,6	1001	90% - 110%
NH	Fall 05	1	3,4,5,6,7,8	835	--
NJ	Spring 05	2	3,4	1123	92.5% - 107.5%
NJ	Spring 06	2	3,4,5,6,7	1599	90% - 110%
NM	Spring 05	1	3,4,5,6,7,8	2758	--
NM	Spring 06	2	3,4,5,6,7,8	3740	95% - 105%
NV	Spring 03	2	3,5	1275	95% - 105%
NV	Spring 06	1	3,4,5,6.7,8	979	--
OH	Spring 07	2	3,4,5,6,7,8	1352	92.5% - 107.5%
RI	Fall 05	From New Hampshire results	--	--	--
SC	Spring 02	1	3,4,5,6,7,8	1931	--
SC	Spring 06	2	3,4,5,6,7,8	20414	95% - 105%
TX	Spring 03	1	5,7	3252	--
TX	Spring 06	2	3,4,5,6,7	2435	95% - 105%
VT	Fall 05	From New Hampshire results	--	--	
WA	Spring 04	1	4,7,10	4248	--
WA	Spring 06	2	3,4,5,6,7,8	14825	95% - 105%
WI	Fall 03	1	4,8	724	--
WI	Fall 05	2	3,4,5,6,7,8	5327	--

Note: Method 1 = Direct Estimate; Method 2 = Indirect Method

Table A3.2 – Summary of Study Method and Sample Population by State - **Reading**

State	Term	Method	Grades	Average student count per grade	Match Level
AZ	Spring 02	1	3, 5, 8	2368	--
AZ	Spring 05	1	3,4,5,6,7,8	2828	--
CA	Spring 03	1	3,4,5,6,7,8	10446	--
CA	Spring 06	2	3,4,5,6,7,8	7353	95% - 105%
CO	Spring 02	1	4,5,6,7,8	5643	--
CO	Spring 05	1	3,4,5,6,7,8	3318	--
DE	Spring 06	1	3,4,5,6,7,8	1914	--
ID	Spring 03	NWEA administered state test	3,4,5,6,7,8	--	--
ID	Spring 06	NWEA administered state test	3,4,5,6,7,8	--	--
IL	Spring 03	1	3,5,7,8	1499	--
IL	Spring 06	1	3,4,5,6,7,8	1223	--
IN	Fall 02	1	3,6,8	2683	--
IN	Fall 06	2	3,4,5,6,7,8	13610	95% - 105%
KS	Fall 06	1	3,4,5,6,7,8	2269	--
MA	Spring 06	2	3,4,5,6,7,8	1591	92.5% - 107.5%
MD	Spring 05	1	3,4,5	8188	--
MD	Spring 06	2	3,4,5,6,7,8	8145	95% - 105%
ME	Spring 06	2	3,4,5,6,7,8	1818	95% - 105%
MI	Fall 03	2	4,7	1179	95% - 105%
MI	Fall 05	1	3,4,5,6,7,8	2490	--
MN	Spring 03	1	3,5,8	4366	--
MN	Spring 06	1	3,4,5,6,7,8	12105	--
MT	Spring 04	1	4,8	1465	--
MT	Spring 06	2	3,4,5,6,7,8	1868	95% - 105%
ND	Fall 04	1	3,4,5,6,7,8	1521	--
ND	Fall 06	2	3,4,5,6,7,8	1817	90% - 110%
NH	Fall 03	2	3,6	987	90% - 110%
NH	Fall 05	1	3,4,5,6,7,8	833	--
NJ	Spring 05	2	3,4	986	92.5% - 107.5%
NJ	Spring 06	2	3,4,5,6,7,8	2601	90% - 110%
NM	Spring 05	1	3,4,5,6,7,8	2014	--
NM	Spring 06	2	3,4,5,6,7,8	3323	95% - 105%
NV	Spring 03	2	3,5	1206	95% - 105%
NV	Spring 06	1	3,4,5,6,7,8	1007	--
OH	Spring 07	2	3,4,5,6,7,8	1297	92.5% - 107.5%
RI	Fall 05	From New Hampshire results	--	--	--
SC	Spring 02	1	3,4,5,6,7,8	1932	--
SC	Spring 06	2	3,4,5,6,7,8	18669	95% - 105%
TX	Spring 03	1	3,5	2947	--
TX	Spring 06	2	3,4,5,6,7	2435	95% - 105%
VT	Fall 05	From New Hampshire results	--	--	--
WA	Spring 04	1	4,7	5616	--
WA	Spring 06	2	3,4,5,6,7,8	14794	95% - 105%
WI	Fall 03	1	4,8	725	--
WI	Fall 05	2	3,4,5,6,7,8	4985	95% - 105%

Note: Method 1 = Direct Estimate; Method 2 = Indirect Method

Appendix 4 - Estimated State-Test Proficiency Cut Scores in Reading using MAP (in Percentile Ranks)

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
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
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
Maryland	26	20	23	23	27	31
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	36
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	unavailable
Vermont	33	34	34	43	40	48
Washington	37	23	27	40	49	36
Wisconsin	14	16	16	16	17	14
Median for 26 states	31	29	30	32	32	36

Appendix 5 - Estimated State-Test Proficiency Cut Scores in Mathematics using MAP (in Percentile Ranks)

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Arizona	30	28	33	40	36	42
California	46	55	57	62	59	unavailable
Colorado	6	8	9	16	19	25
Delaware	25	26	24	29	36	36
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	unavailable
New Mexico	46	49	54	60	61	56
North Dakota	20	27	23	32	39	41
Ohio	20	32	40	34	32	32
Rhode Island	41	35	34	44	44	53
South Carolina	71	64	72	65	68	75
Texas	30	34	24	35	41	unavailable
Vermont	41	35	34	44	44	53
Washington	36	46	48	57	59	56
Wisconsin	29	29	26	21	21	23
Median for 25 states	30	35	34	40	43	45

Note: There was not sufficient data to generate eighth grade estimates for California, New Jersey, and Texas.

Appendix 6 – Changes in Proficiency Cut Score Estimates and Reported Proficiency Rates on State Assessments – Reading

		Change in proficiency cut score (in percentile ranks)			Change in state reported proficiency		
State	Grade	Current cut score	Prior cut score	Change	Current proficiency	Prior proficiency	Change
Arizona	Grade 3	23	26	-3	72%	75%	-3%
	Grade 5 *	25	37	-12	71%	59%	12%
	Grade 8 *	36	47	-11	67%	56%	11%
California	Grade 3	61	58	3	36%	33%	3%
	Grade 4 *	43	55	-12	49%	39%	10%
	Grade 5	53	60	-7	43%	36%	7%
	Grade 6	56	59	-3	41%	36%	5%
	Grade 7 *	52	61	-9	43%	36%	7%
	Grade 8 *	56	68	-12	41%	30%	11%
Colorado	Grade 3 *	7	16	-9	90%	90%	0%
	Grade 4 *	11	14	-3	86%	85%	1%
	Grade 5 *	11	15	-4	88%	83%	5%
	Grade 6	13	12	1	87%	86%	1%
	Grade 7	17	18	-1	85%	83%	2%
	Grade 8	14	16	-2	86%	85%	1%
Illinois	Grade 3 *	35	52	-17	71%	62%	9%
	Grade 5	32	35	-3	69%	60%	9%
	Grade 8 *	22	36	-14	79%	64%	15%
Indiana	Grade 3	27	29	-2	73%	72%	1%
	Grade 6	32	29	3	71%	68%	3%
	Grade 8	33	39	-6	67%	63%	4%
Maryland	Grade 3 *	26	33	-7	78%	76%	2%
	Grade 4	20	21	-1	82%	81%	1%
	Grade 5 *	23	32	-9	77%	74%	3%
Michigan	Grade 4	20	19	1	83%	75%	8%
	Grade 7 *	25	37	-12	76%	61%	15%
Minnesota	Grade 3 *	26	33	-7	82%	76%	6%
	Grade 5	32	27	5	77%	81%	-4%
	Grade 8 *	44	36	8	65%	81%	-16%
Montana	Grade 4 *	25	37	-12	80%	66%	14%
	Grade 8 *	36	53	-17	76%	58%	18%
Nevada	Grade 3 *	46	55	-9	51%	48%	3%
	Grade 5	53	57	-4	39%	46%	-7%

Appendix 6 – Continued

State	Grade	Change in proficiency cut score (in percentile ranks)			Change in state reported proficiency		
		Current cut score	Prior cut score	Change	Current proficiency	Prior proficiency	Change
New Hampshire	Grade 3 *	33	18	↑ 15	71%	75%	↓ -4%
	Grade 6 *	43	30	↑ 13	65%	74%	↓ -9%
New Jersey	Grade 3 *	15	12	↑ 3	82%	83%	↓ -1%
	Grade 4 *	25	17	↑ 8	80%	82%	↓ -2%
New Mexico	Grade 3	33	33	→ 0	55%	55%	↑ 0%
	Grade 4	32	34	↓ -2	54%	52%	↑ 2%
	Grade 5	30	30	→ 0	57%	57%	→ 0%
	Grade 6	43	43	→ 0	40%	41%	↓ -1%
	Grade 7	32	35	↓ -3	50%	50%	→ 0%
	Grade 8	33	39	↓ -6	51%	52%	↓ -1%
North Dakota	Grade 3 *	22	33	↓ -11	78%	78%	→ 0%
	Grade 4	29	34	↓ -5	78%	82%	↓ -4%
	Grade 5	34	37	↓ -3	73%	78%	↓ -5%
	Grade 6	37	34	↑ 3	72%	79%	↓ -7%
	Grade 7	30	34	↓ -4	76%	79%	↓ -3%
	Grade 8	33	36	↓ -3	69%	74%	↓ -5%
South Carolina	Grade 3 *	43	61	↓ -18	55%	42%	↑ 13%
	Grade 4 *	58	68	↓ -10	42%	34%	↑ 8%
	Grade 5 *	64	76	↓ -12	34%	25%	↑ 9%
	Grade 6	62	65	↓ -3	31%	34%	↓ -3%
	Grade 7	69	72	↓ -3	26%	27%	↓ -1%
	Grade 8	71	71	→ 0	25%	27%	↓ -2%
Texas	Grade 3 *	12	6	↑ 6	89%	85%	↑ 4%
	Grade 5 *	30	19	↑ 11	80%	79%	↑ 1%
	Grade 6 *	21	16	↑ 5	91%	86%	↑ 5%
	Grade 7 *	32	20	↑ 12	79%	87%	↓ -8%
Washington	Grade 4 *	23	29	↓ -6	81%	74%	↑ 7%
	Grade 7	49	49	→ 0	62%	60%	↑ 2%
Wisconsin	Grade 4	16	15	↑ 1	82%	81%	↑ 1%
	Grade 8 *	14	20	↓ -6	85%	79%	↑ 6%

* Indicates that the change was greater than one standard error of measure on MAP

Appendix 7 – Changes in Proficiency Cut Score Estimates and Reported Proficiency Rates on State Assessments - Mathematics

State	Grade	Change in proficiency cut score (in percentile ranks)			Change in state reported proficiency		
		Current cut score	Prior cut score	Change	Current proficiency	Prior proficiency	Change
Arizona	Grade 3 *	30	39	-9	77%	62%	15%
	Grade 5 *	33	51	-18	71%	46%	25%
	Grade 8 *	42	78	-36	63%	21%	42%
California	Grade 3	46	50	-4	58%	46%	12%
	Grade 4	55	52	3	54%	45%	9%
	Grade 5 *	57	65	-8	48%	35%	13%
	Grade 6	62	62	0	41%	34%	7%
	Grade 7 *	59	72	-13	41%	30%	11%
Colorado	Grade 5 *	9	13	-4	89%	86%	3%
	Grade 6	16	16	0	85%	81%	4%
	Grade 7 *	19	24	-5	82%	75%	7%
	Grade 8 *	25	31	-6	75%	70%	5%
Illinois	Grade 3	20	22	-2	86%	76%	10%
	Grade 5 *	20	28	-8	79%	68%	10%
	Grade 8 *	20	47	-27	78%	53%	25%
Indiana	Grade 3	35	41	-6	72%	67%	5%
	Grade 6 *	27	36	-9	80%	68%	12%
	Grade 8	34	36	-2	71%	66%	5%
Michigan	Grade 4 *	13	18	-5	82%	65%	17%
	Grade 8	32	30	2	63%	52%	11%
Minnesota	Grade 3	30	36	-6	78%	75%	3%
	Grade 5 *	54	26	28	59%	77%	-18%
	Grade 8 *	51	44	7	57%	72%	-15%
Montana	Grade 4 *	43	55	-12	64%	45%	19%
	Grade 8 *	60	44	16	58%	64%	-7%
Nevada	Grade 3	50	50	0	51%	50%	1%
	Grade 5	46	46	0	45%	50%	-5%
New Hampshire	Grade 3 *	41	6	35	68%	84%	-16%
	Grade 6 *	44	22	22	61%	73%	-12%

Appendix 7 – Continued

State	Grade	Change in proficiency cut score (in percentile ranks)			Change in state reported proficiency		
		Current cut score	Prior cut score	Change	Current proficiency	Prior proficiency	Change
New Jersey	Grade 3 *	13	22	↓ -9	87%	83%	↑ 4%
	Grade 4	23	28	↓ -5	82%	80%	↑ 2%
New Mexico	Grade 3	46	46	→ 0	45%	43%	↑ 2%
	Grade 4	49	49	→ 0	41%	39%	↑ 2%
	Grade 5	54	60	↓ -6	34%	27%	↑ 7%
	Grade 6 *	60	67	↓ -7	24%	22%	↑ 2%
	Grade 7	61	66	↓ -5	23%	20%	↑ 3%
	Grade 8 *	56	62	↓ -6	26%	24%	↑ 2%
North Dakota	Grade 3	20	22	↓ -2	85%	87%	↑ 2%
	Grade 4	27	27	→ 0	78%	84%	↓ -6%
	Grade 5 *	23	34	↓ -11	78%	78%	→ 0%
	Grade 6	32	36	↓ -4	76%	78%	↓ -2%
	Grade 7	39	37	↑ 2	71%	74%	↓ -3%
	Grade 8	41	43	↓ -2	66%	67%	↓ -1%
South Carolina	Grade 3	71	64	↑ 7	35%	32%	↑ 3%
	Grade 4	64	64	→ 0	42%	36%	↑ 6%
	Grade 5	72	75	↓ -3	34%	29%	↑ 5%
	Grade 6 *	65	72	↓ -7	37%	29%	↑ 8%
	Grade 7	68	72	↓ -4	32%	27%	↑ 5%
	Grade 8 *	75	80	↓ -5	22%	19%	↑ 3%
Texas	Grade 5 *	24	13	↑ 11	81%	86%	↓ -5%
	Grade 7 *	41	25	↑ 16	70%	73%	↓ -3%
Washington	Grade 4	46	49	↓ -3	59%	60%	↓ -1%
	Grade 7	59	61	↓ -2	49%	46%	↑ 2%
Wisconsin	Grade 4	29	27	↑ 2	73%	73%	→ 0%
	Grade 8 *	23	34	↓ -11	74%	65%	↑ 9%

* Indicates that the change was greater than one standard error of measure on MAP

Appendix 8 - How Consistent Are the Results from this Study and the NCES Mapping 2005 State Proficiency Standards Study?

A number of prior studies have attempted to compare the difficulty of proficiency standards across states, the most recent being a report published by the National Center for Educational Statistics (2007) that estimated thirty-three state proficiency cut scores using data from the 2005 National Assessment of Educational Progress. We wanted to know whether our results were consistent with those of the NCES.

We started by comparing the two studies' individual estimates of cut scores by state. NAEP reading and math assessments are administered to students in grades 4 and 8. For fourth grade, we found sixteen states with estimates of cut scores derived from MAP as well as NAEP in both reading and math. For eighth-grade, we found fifteen states with estimates from both MAP and NAEP in reading, and thirteen states with estimates from both in mathematics. The NAEP cut score estimates were computed using data from the spring 2005 testing season, while the MAP cut score estimates were computed using the most recent available testing data – either the 2005, 2006, or 2007 testing seasons.

Estimates of cut scores derived from NAEP were generally consistent with estimates derived from MAP.

In order to correlate the estimated cut scores from the two studies, we converted the cut score estimates from each study to rank scores, and calculated Spearman's Rho (an indicator that measures the degree of correlation between ranked variables) on the matched pairs of ranks (see Table A8.1). The results show moderate correlations between NCES rankings and those reported in this study, suggesting that the rankings produced by the two studies are similar but not identical. In order to evaluate the magnitude of differences between the two sets of estimates, we also converted the scale score estimates for both studies to z scores (a simple metric for comparing scores from different scales) and calculated the differences. Figures A8.1 through A8.4 show the results of those analyses.

Table A8.1 – Spearman's Rho correlation of NAEP and MAP estimates of proficiency cut scores based on ranking of difficulty

	States evaluated	Spearman's Rho
Grade 4 – Reading	16	.63
Grade 4 – Mathematics	16	.65
Grade 8 – Reading	15	.63
Grade 8 – Mathematics	13	.62

Figure A8.1 - Z score differences between NAEP and MAP estimated proficiency cut scores in grade 4 reading

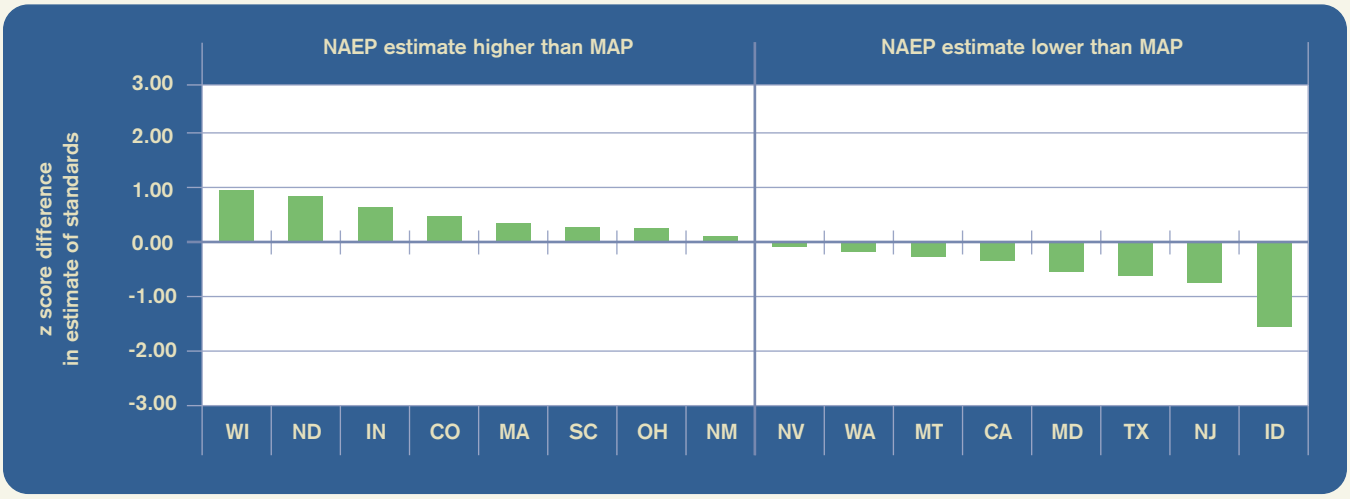


Figure A8.2 - Z score differences between NAEP and MAP estimated proficiency cut scores in grade 4 mathematics

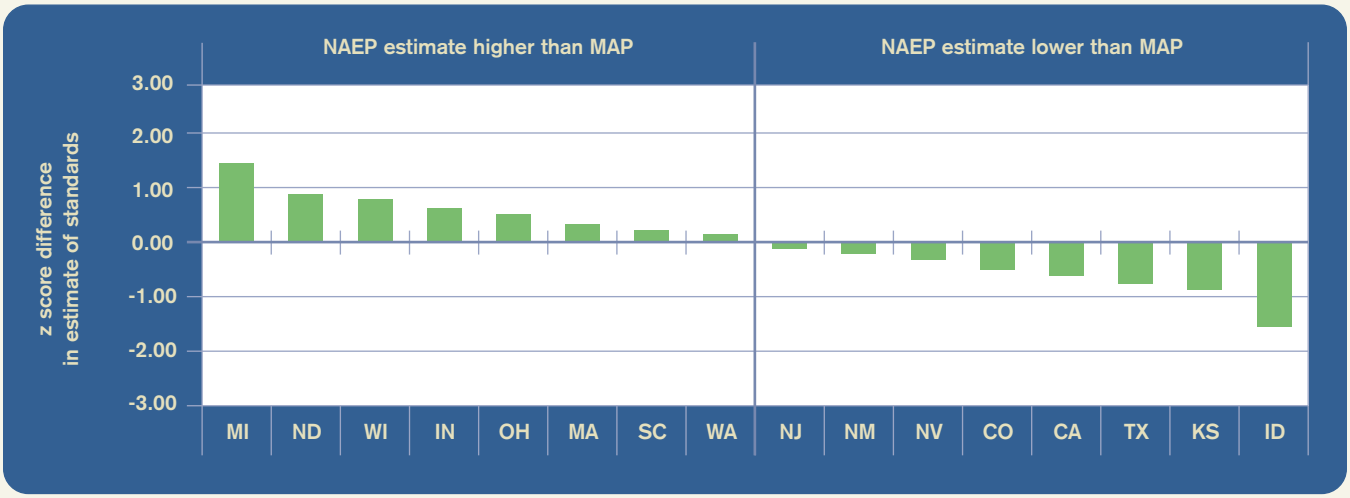


Figure A8.3 - Z score differences between NAEP and MAP estimated proficiency cut scores in grade 8 reading

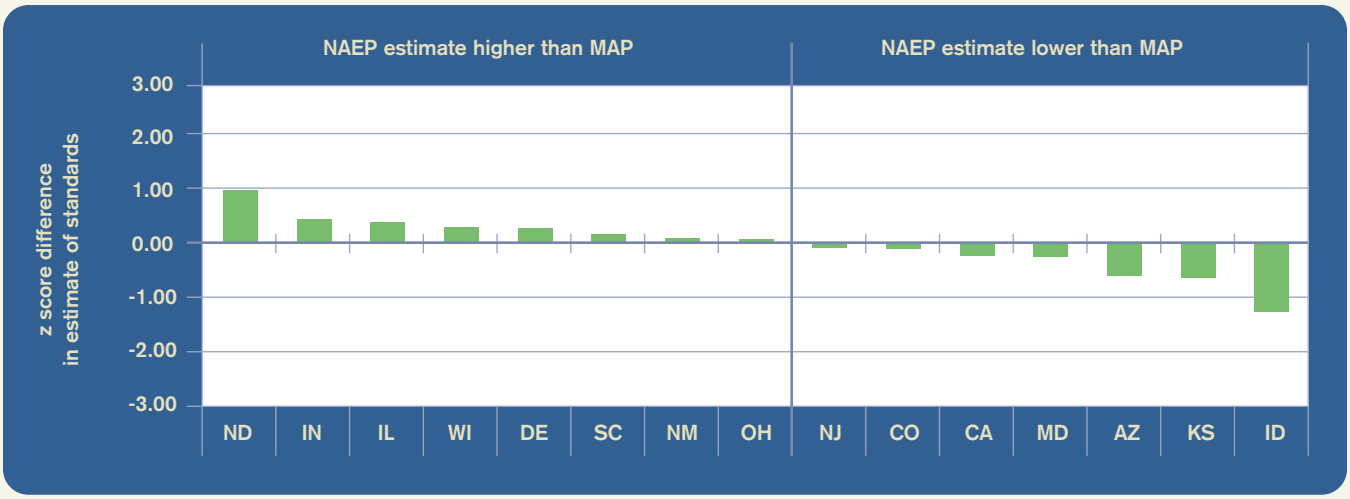
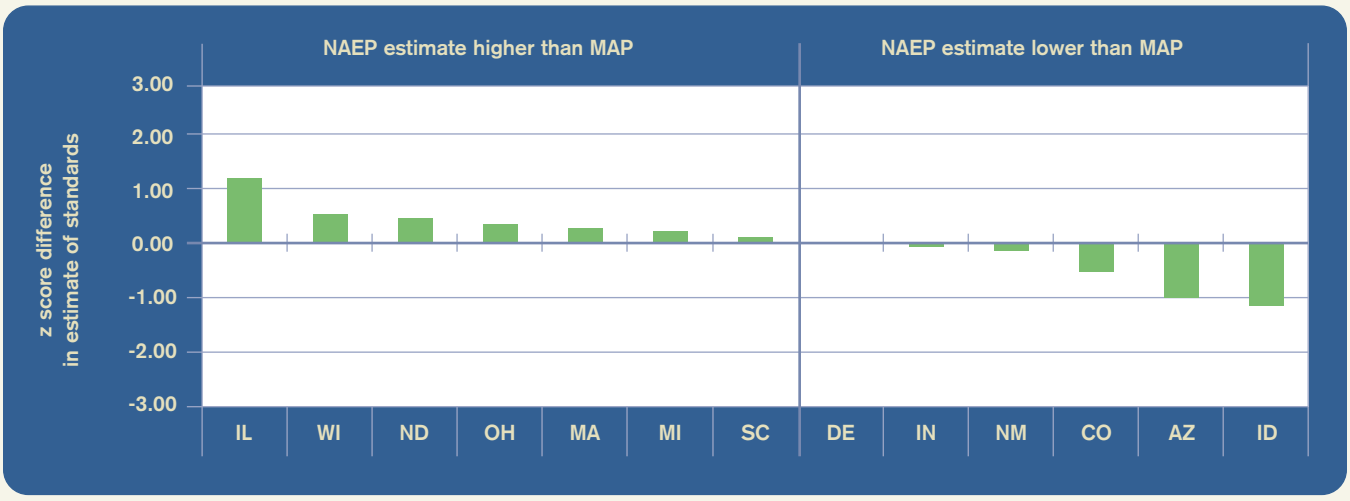


Figure A8.4 - Z score differences between NAEP and MAP estimated proficiency cut scores in grade 8 mathematics



Figures A8.1 - A8.4 show that the majority of standardized cut score estimates were within 0.5 z across grades and subjects. There were several exceptions. For example, several of the states for which the NAEP estimates were higher than MAP estimates by more than 0.5 z were those that administer their test during the fall season, including Michigan, North Dakota, Wisconsin, and Indiana. The MAP scores used to generate proficiency cut scores estimates were collected during the same season in which the state test was administered. Thus, when the state test is administered in the autumn, the MAP estimate is based on the fall test. NAEP, however, is administered only in spring, so the NAEP estimate of the cut scores for these fall tests is based on a spring result. Because students in these states will have had additional instructional time and opportunity for growth between fall and spring, their NAEP score will reflect as much. Thus, the NAEP estimate of the cut score in these states is likely to be slightly higher than the MAP estimate. This effect is reflected in the data, where states engaged in fall testing show consistently higher NAEP estimates than MAP estimates. Had the NCES study been able to control for this time difference, the estimates would very likely have been even closer than those reported.

NWEA also provided the state test for Idaho during this period, and the NAEP estimate of the cut score was much lower, on a relative basis, than our own. This may illustrate a point made earlier in this report, that some outside factors lead to increases in performance on the NWEA test that are not reflected in NAEP. As a result, it is possible that student performance gains in Idaho on MAP would not have been entirely replicated on NAEP.

Both studies found that math cut scores were generally higher than reading cut scores.

As noted above, according to MAP estimates, state proficiency standards in mathematics were generally more difficult than those in reading. This analysis used normative conversions of scale score data to evaluate the difficulty of standards. Thus, if a state's reading cut score for fourth grade is set at a scale score equivalent to the 40th percentile and its math cut score is at the 60th, we can fairly say the mathematics standard is more difficult. NAEP, however, is not normed, so we used the means and standard deviations reported for the 2005 administration of NAEP to estimate z values for the NCES study's cut score estimates. Averaging these z values and returning their percentile rank in a normal distribution provided one way of estimating the difficulty of the fourth- and eighth-grade cut score estimates across the states studied.

The NCES study included twenty-seven states that had both fourth- and eighth-grade estimates for reading and twenty-nine states that had both estimates for mathematics. The NCES results (Table A8.2) show small differences in the difficulty of math and reading standards at fourth grade, with mathematics cut scores being approximately 4 percentile ranks more difficult. In eighth grade, however, the difference was considerably larger: the math cut scores were the equivalent of 10 percentile ranks more difficult than the reading cut scores. Both results are consistent with our analyses, which found mathematics cut scores set at more challenging levels than reading cut scores in all grades, with larger differences found in the upper grades.

Table A8.2 – Differences in NCES reading and mathematics cut score estimates by grade

GRADE 4				GRADE 8			
Reading		Mathematics		Reading		Mathematics	
z	Percentile rank	z	Percentile rank	z	Percentile rank	z	Percentile rank
-.65	26	-.52	30	-.47	32	-.21	42

Both studies found that cut scores decreased more than they increased across the time periods studied, excepting those for grade 4 mathematics.

The NCES study focused on its 2005 estimates of state proficiency cut scores, but the study also reported 2003 state proficiency estimates in an appendix. The authors note that the results of the two analyses may not be comparable because of changes in relevant state policies that may have occurred during the study period. However, because our study was interested in whatever changes may have occurred in the standards, regardless of why they occurred, we summarized the data in the NCES estimates in an effort to see if the data showed similar direction in the perceived changes in standards.

Because the NCES study used NAEP data, comparisons were limited to grades 4 and 8. In addition, many of the states studied by NCES differed from ours, and the cut score estimates were not always generated at the same time. As a result, we did not directly compare changes in particular state estimates between the two studies. Table A8.3 summarizes the differences in the NCES estimates between 2003 and 2005. These show that cut score estimates decreased more than they increased in fourth-grade reading, as well as in eighth-grade reading and math. In fourth-grade math, the number of cut score estimate increases was the same as the number of decreases. Everywhere else, the NCES results are consistent in direction with our own.

Table A8.3 – Difference between 2003 and 2005 NCES estimates of state proficiency cut scores using NAEP

	READING		MATHEMATICS	
	Grade 4	Grade 8	Grade 4	Grade 8
States studied	24	28	25	32
Increase	6 (25.0%)	6 (21.4%)	11 (44.0%)	6 (18.8%)
No change	1 (4.1%)	3 (10.7%)	3 (12.0%)	5 (15.6%)
Decrease	17(70.8%)	19 (67.8%)	11(44.0%)	21(65.6%)

Both studies found evidence that reading and math cut scores were not calibrated between grades 4 and 8.

The same methods used to compare the relative difficulties of reading and math cut scores can be utilized to compare the calibration of each subject's cut scores across grades. Because the MAP test is normed, one can evaluate the difficulty of standards between grades by using percentile ranks. Thus, as explained above, if the fourth-grade standard is set at the 40th percentile and the eighth-grade standard is at the 60th, we can fairly say the standards are not calibrated. As in the earlier analysis, we compensated for the fact that NAEP is not normed by using the means and standard deviations reported for the 2005 administration of NAEP to estimate z values for the NCES study's cut score estimates. By averaging these z values and returning their percentile position in a normal distribution, we were able to compare the difficulty of fourth- and eighth-grade cut score estimates across the states studied.

Table A8.4 shows the z values and percentile ranks associated with the average of the cut score estimates. In both subjects, the eighth-grade standards were, on average, more difficult than the fourth-grade standards, with the difference being larger in math (.32 z and 12 percentile ranks) than in reading (.18 z and 6 percentile ranks). The nature and direction of the differences were consistent with our study, which found that grade 8 cut scores were generally more challenging than those of earlier grades, and that the differences were somewhat larger in mathematics than in reading.

In general, the findings of the two studies appear consistent. Both found considerable disparity in the difficulty of standards across states. For states in which both studies estimated cut scores, we found moderate correlations between the rankings by difficulty; many of the differences in ranking can be attributed to the fact that we used fall MAP data to estimate the cut scores for some states while NAEP was limited to using its own spring administrations. Data from both studies support the conclusion that mathematics cut scores are generally set at more difficult levels than reading cut scores. Data from both studies also support the conclusion that state proficiency cut scores have declined more often than they have increased in the period between their respective estimates. Finally, data from both studies support the conclusion that cut scores for students in the upper grades are generally more difficult than in the lower grades.

Table A8.4 – NCES study's estimate of the distribution of state proficiency cut scores estimates

READING				MATHEMATICS			
Grade 4		Grade 8		Grade 4		Grade 8	
z	Percentile rank	z	Percentile rank	z	Percentile rank	z	Percentile rank
-.65	26	-.47	32	-.52	30	-.21	42

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, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

- Anatasi, A., and S. Urbina. 1997. *Psychological Testing.* 7th ed. New York: MacMillan.

- Association of Test Publishers. 2000. *Guidelines for Computer-Based Testing.* Washington, DC: Association of Test Publishers.

- Booher-Jennings, J. 2005. Below the bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal* 42 (2): 231-268.

- Braun, H. 2004. Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives* 12 (1), http://epaa.asu.edu/epaa/v12n1/ (accessed September 8, 2007).

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

- Carnoy, M., and S. Loeb 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis* 24 (4): 305-331.

- Cronin, J., G. G. Kingsbury, M. McCall, and B. Bowe (2005). The Impact of the No Child Left Behind Act on Student Achievement and Growth: 2005 Edition. Lake Oswego, OR: Northwest Evaluation Association.

- Cronin, J. 2006. The effect of test stakes on growth, response accuracy, and item-response time as measured on a computer-adaptive test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

- Cronin, J., G. G. Kingsbury, M. Dahlin, D. Adkins, and B. Bowe. 2007. 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.

- Education Trust. 2004. Measured progress: Achievement rises and gaps narrow but too slowly. Washington, DC: Education Trust, http://www2.edtrust.org/edtrust/images/MeasuredProgress.doc.pdf (accessed September 10, 2007).

- Educational Testing Service. 1991. *The Results of the NAEP 1991 Field Test for the 1992 National and Trial State Assessments.* Princeton, NJ: Educational Testing Service.

References (continued)

- Fuller, B., J. Wright, K. Gesicki, and E. Kang. 2007. Gauging growth: How to judge No Child Left Behind? *Educational Researcher* 36 (5): 268-278.

- Ingebo, G. 1997. *Probability in the Measure of Achievement.* Chicago: Mesa Press.

- Jacob, B. 2002. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. Working paper W8968, National Bureau of Economic Research, Cambridge, MA.

- Kingsbury, G. G. 2003. A long-term study of the stability of item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

- Kingsbury, G. G., A. Olson, J. Cronin, C. Hauser, and R. Houser. 2003. *The State of State Standards: Research Investigating Proficiency Levels in Fourteen States.* Lake Oswego, OR: Northwest Evaluation Association.

- Koretz, Daniel. 2005. Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education* 104 (2): 99–118.

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

- 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. 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.

- McGlaughlin, D. and V. Bandeira de Mello. 2003. Comparing state reading and math performance standards using NAEP. Paper presented at National Conference on Large-Scale Assessment, San Antonio, TX.

- Mullis, I.V.S., M. O. Martin, E. J. Gonzales, and A. M. Kennedy. 2003. *PIRLS 2001 International Report.* Boston: International Study Center.

- Mullis, I.V.S., M. O. Martin, E. J. Gonzales, and S. J. Chrostowski. 2004. *TIMSS 2003 International Mathematics Report.* Boston: International Study Center.

- National Center for Educational Statistics. 2007. *Mapping 2005 State Proficiency Standards onto the NAEP Scales* (NCES 2007-482). Washington: DC: U.S. Department of Education.

References (continued)

- Neal, D. and D. Whitmore-Schanzenbach. 2007. *Left Behind by Design: Proficiency Counts and Test-Based Accountability,* http://www.aei.org/docLib/20070716_NealSchanzenbachPaper.pdf (accessed August 18, 2007).

- New American Media. 2006. Great Expectations: *Multilingual Poll of Latino, Asian and African American Parents Reveals High Educational Aspirations for their Children and Strong Support for Early Education.* San Francisco, CA: New American Media.

- Northwest Evaluation Association (2003, September). Technical Manual for the NWEA Measures of Academic Progress and Achievement Level Tests. Lake Oswego, OR: Northwest Evaluation Association

- Northwest Evaluation Association. 2005a. *Validity Evidence for Achievement Level Tests and Measures of Academic Progress.* Lake Oswego, OR: Northwest Evaluation Association.

- Northwest Evaluation Association. 2005b. *RIT Scale Norms.* Lake Oswego, OR: Northwest Evaluation Association.

- Northwest Evaluation Association. 2007. *Content Alignment Guidelines.* Lake Oswego, OR: Northwest Evaluation Association.

- O'Neil, H., B. Sugrue, J. Abedi, E. Baker, and S. Golan. 1997. *Final Report on Experimental Studies of Motivation and NAEP Test Performance.* CSE Technical Report 427. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

- Rosenshine, B. 2003. High-stakes testing: Another analysis. *Education Policy Analysis Archives* 11 (24), http://epaa.asu.edu.epaa/v11n24 (accessed September 8, 2007).

- Triplett, S. 1995. Memorandum to North Carolina LEA Superintendents. Raleigh, NC: Department of Education, June 11.

- United States Department of Education. 2005. Idaho Assessment Letter, http://www.ed.gov/admins/lead/account/nclbfinalassess/id.html (accessed July 31, 2007).

- White, Katie Weits and James E. Rosenbaum. 2007. Inside the blackbox of accountabilty: How high-stakes accountability alters school culture and the classification and treatment of students and teachers. In *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on Federal Education Policy*, A. Sadvonik, J. O'Day, G. Bohrnstedt, and K. Borman, eds. New York: Routledge.

- Williams, V. S. L., K . R. Rosa, L. D. McLeod, D. Thissen, and E. Sanford. 1998. Projecting to the NAEP scale: Results from the North Carolina End-of-Grade Testing System. *Journal of Educational Measurement* 35: 277-296.

- Wright, B. D. 1977. Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 14 (2): 97-116.

Copies of this report are available electronically
at our website, www.edexcellence.net

Thomas B. Fordham Institute
1701 K Street, N.W.
Suite 1000
Washington, D.C. 20006