

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 equipercntile 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?