KINGSBURY
CENTER AT NWEA

The State of Proficiency:

How student proficiency rates vary across states, subjects, and grades between 2002 and 2010

Sarah Durant and Michael Dahlin

July 2011

Sarah Durant, M.P.P. in Educational Public Policy, is a Senior Research Associate in the Kingsbury Center at NWEA. Her primary research interests are educational equity and making accurate education data available to a variety of stakeholders.

Michael Dahlin, Ph.D. in Developmental Psychology, is a Research Specialist in the Kingsbury Center at NWEA. His primary research interests are policy research related to testing standards, school accountability, and teacher accountability.

John Cronin, Ph.D. in Educational Studies, is the Director of the Kingsbury Center at NWEA. His research interests are wide-ranging, and his recent work has focused on the impact of the No Child Left Behind Act on state standards, equity and the measurement of student growth.

Table of Contents

Table of Figures

EXECUTIVE SUMMARY: The State of Proficiency

Background

In 2007, NWEA and the Thomas B. Fordham Institute collaborated on *The Proficiency Illusion*, a study that illustrated the issues created by having each state set its own standards for what constitutes student proficiency for reading and mathematics tests. By comparing the cut scores that determine proficiency for each state, we found that there was significant variation in the difficulty of proficiency levels among states.

In the four years since the study was published, the educational landscape has obviously changed. At the time of this printing, the Congress was still considering the reauthorization of the Elementary and Secondary Education Act (ESEA). It is likely that many of the accountability measures inside No Child Left Behind (NCLB) may change as part of the reauthorization. And the current genre of proficiency tests are likely to eventually be replaced by tests developed by the Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium (SBAC). Among the goals of the consortia is the development of tests that measure student performance against standards of college readiness and the establishment of consistent benchmarks for performance across states.

As their work proceeds, it seems important to look at the current state of proficiency standards relative to the expectations these new assessments are likely to establish. So we updated our original study with recent data and enhanced the visualizations showing how states, subjects, and grades compare in terms of proficiency. NWEA is in a unique position to investigate this question because we have a large sample of data collected from schools whose students participated both in state testing and in the NWEA Measures of Academic Progress (MAP) assessment, allowing us to use the NWEA scale as a common ruler for students in multiple grades, subjects, and states. We hope that the information in this study can help inform the next generation of policies governing our nation's schools.

Findings

The findings section explores several main concerns around using proficiency data to make educational decisions:

- Proficiency standards vary across states, and in nearly all states studied, they remain far below any level that would be characterized as college readiness.
- Most of the differences in proficiency rates that are seen across states, and across grades within states, are a function of the difficulty of the state tests, not differences in student performance.
- Because standards remain uncalibrated across grades and subjects, proficiency cut scores in the upper grades are frequently more demanding, sometimes far more demanding, than cut scores in the early grades. Math cut scores are also frequently higher than reading cut scores.
- In general, the difficulty of proficiency standards within states has not changed dramatically over time. However, when they have changed, they have grown harder about as often as they have grown easier.

These issues are likely to create significant challenges for educators as we move from the current genre of state tests to assessments that will evaluate students against considerably higher standards of performance. Some of the likely problems we anticipate are these:

1. Educators who have aligned their curricula and instruction to standards that average between the 30th and 40th percentile in difficulty will need to make major changes in their classrooms if they are to deliver curricula that are aligned to true benchmarks of college readiness.
2. Because proficiency standards have typically been considerably easier in the early grades than upper grades, teachers in the early elementary grades in particular may be especially challenged if the new assessments have cut scores that accurately reflect the level of achievement required to be college ready.
3. The current lack of calibration of proficiency cut scores across grades continues to communicate misinformation about the performance of schools. In particular, it contributes to a myth that middle schools are less effective in schooling than elementary schools. Such misinformation currently results in far more middle schools than elementary schools being identified for NCLB sanctions, creating potential misallocation of precious and increasingly scarce educational resources.
4. Students and parents want a clear college and career trajectory for K-12. The current genre of tests has not provided this. While we are encouraged by the goals of the assessment consortia, it is important that both the curricula supported by the standards and the cut scores associated with the assessment reflect true college readiness standards. In addition, it is particularly important that the standards are calibrated in a manner that assures that students who are deemed college ready in the early grades are truly destined to be college ready in upper grades and high school, assuming they make normal progress.

Recommendations

Based on the findings from this study and other research, we make the following recommendations for policymakers as they reauthorize the Elementary and Secondary Education Act:

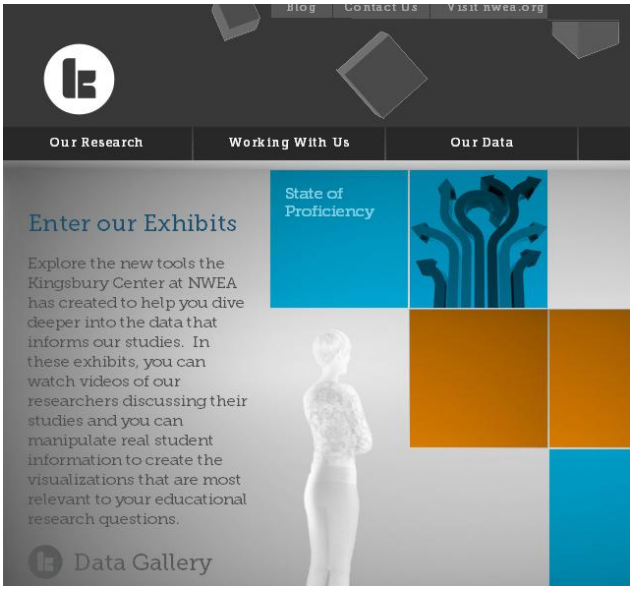
- We recommend that new standards and assessment systems should be structured based on what students should know and be able to do at the end of high school, and the proficiency/mastery standards at each grade scaled accordingly so students know where they are in meeting that target.

- We recommend that assessment systems should be scaled and calibrated to reflect equivalent levels of difficulty across subjects or that the intention of scaling subjects differently is clearly articulated and understood.

- We recommend that the new assessment systems either use an equivalent measurement scale or that a reliable crosswalk between systems is readily available so students can be compared, both across states and across grades.

Data Gallery

This report is only one part of the work we have done to portray the differences in state standards and how these differences can result in misinterpretation of student performance or misallocation of resources. Because we believe that it is important for educational advocates to understand how the data function in different uses, we have also developed an online data gallery that allows users to interact with real data to see for themselves the effects of different types of policies.



The data galleries can be accessed at http://www.kingsburycenter.org/gallery.

Each data exhibit includes video clips, interactive data visualizations using real data from the study, and links to other studies and blog posts that are related to the study topic. There is also space for visitors to leave comments, ask questions, and share.

We hope that this study and the interactive data galleries will help inform the important discussions happening in the field of education right now and will enhance the ability of education officials to improve the educational system so that all kids can learn.

INTRODUCTION: Introducing the State of Proficiency

In 2007, NWEA and the Thomas B. Fordham Institute collaborated on *The Proficiency Illusion*, a study that illustrated the issues created by having each state set its own standards for what constitutes student proficiency for reading and mathematics tests, while holding all states to the same accountability standards. By comparing the cut scores that determine proficiency for each state, we found that there was significant variation in the difficulty of proficiency levels among states. In some states, it is considerably easier for students to pass their state tests than it is for students in other states.

In the four years since the study was published, the educational landscape has changed in many ways. For instance, the No Child Left Behind (NCLB) legislation required states to achieve 100% proficiency for students by the year 2014. As the deadline draws closer, most states fall far short of reaching that goal, creating an incentive for states to lower their standards.[1] The political administration that inherited NCLB in 2008 has recognized this issue along with other challenges related to the NCLB legislation and is working to address some of these issues in the reauthorization of the Elementary and Secondary Education Act (ESEA).

Another major change since the publication of *The Proficiency Illusion* is the increasingly widespread advocacy by educators and policymakers for shared content standards among states. The National Governors' Association, in collaboration with the Chief Council of State School Officers, created a set of common core curriculum standards for use throughout the country,[2] and several states have already adopted these standards. Concurrently, it is recognized that new assessments will be needed to measure student learning in relation to these standards. As states affiliate themselves with one of the two newly formed assessment consortia that will be developing new systems to measure student proficiency and progress, discussions are happening across the country about how to maintain local control over education while still ensuring rigorous national expectations.[3]

Yet another major change underway across the country is the pressure from educational officials and policymakers to measure the effectiveness of teachers using student assessment data. Federal grant programs such as Race to the Top have required using student data in teacher evaluations, and many states have worked with teacher unions to implement systems to use student data to measure teacher effectiveness.[4] As schools and states begin to design and implement their evaluation programs, important decisions about performance pay, promotion, tenure, and dismissal are being made based on underlying data that was never intended for such uses. The application of such data based on inconsistent state-defined proficiency levels means that these evaluation programs may not be producing the desired effect.

---

[1] Knepper, Matthew D. (2009). Teaching Federal Courts: The Innocence of the No Child Left Behind Act's One Hundred Percent Proficiency Goal and its Consequences. *Saint Louis University Law Journal*. 53(3). Retrieved from http://heinonline.org/HOL/LandingPage?collection=journals&handle=hein.journals/stlulj53&div=31&id=&page=

[2] http://www.corestandards.org/

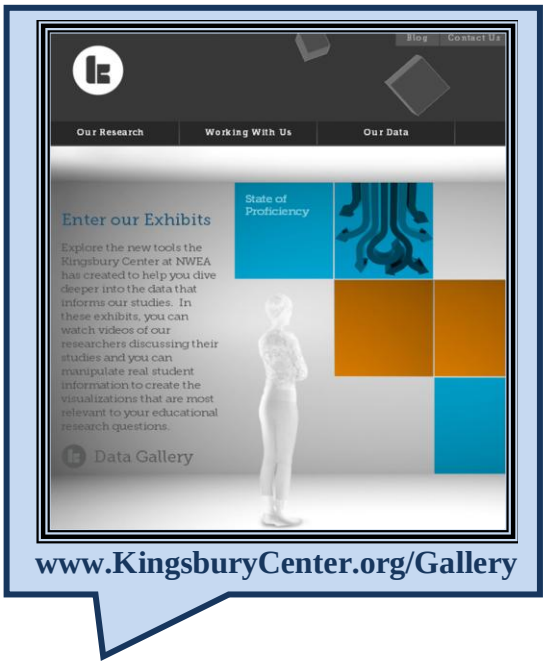
[3] Gewertz, Catherine. (2010) Critics Post 'Manifesto' Opposing Shared Curriculum. *Education Week*. May 10, 2011. Retrieved from http://www.edweek.org/ew/articles/2011/05/09/31curriculum.h30.html

[4] Rotherham, Andrew J. (2010) Rating Teachers: The Trouble with Value-Added Data. *Time Magazine*. September 23, 2010. Retrieved from http://www.time.com/time/nation/article/0,8599,2020867,00.html

Given all of the changes taking place in the field of education, and fortified with more recent data, we decided to update and enhance the original study so that it might inform the next generation of policies governing our nation's schools. In the last decade, the term "proficiency rate" has entered the mainstream lexicon as a measure of school quality, with most people having at least an intuitive understanding that proficiency rates are defined as the number of students who pass the state test, divided by the number of students taking the test. What may be less understood by the general public, however, is that "proficiency" has no objective meaning; it is largely determined by the choices a state makes in creating its assessment standards, and is not connected to any external criteria (such as college readiness) that are independent of the test. The purpose of this study is to shine some light on the limitations of using proficiency rates based on inconsistent and arbitrary "passing scores" to make judgments about educational effectiveness.

This report serves as a written summary of our methodology and findings, but we also believe that it is important for educational advocates to understand for themselves how the data function when used for various policy purposes. For this reason, we have created an online, interactive data gallery where users can explore different states, subjects, and grades to see how proficiency rates change under different circumstances.

The data galleries can be accessed at http://www.kingsburycenter.org/gallery. This report makes frequent reference to the galleries in call-out boxes such as the one to the right. We hope that this study and the interactive data galleries will help inform the important discussions happening in the field of education right now and will enhance the ability of education officials to improve the educational system so that all kids can learn.



METHODOLOGY: How the State of Proficiency was calculated

Cut scores for the various state tests were expressed on a single common scale so that direct comparisons of the difficulty of proficiency standards could be made. State test proficiency standards were linked to the scale of NWEA's Measures of Academic Progress (MAP), a computerized adaptive test of academic achievement used by more than 4,500 school systems across all 50 states as well as in over 100 countries internationally. Although the specific items seen by students taking MAP assessments are aligned to state content standards within each state, the items are all linked to a single common scale, making it possible to directly compare MAP scores across different states and grades.

The assessment is designed to be adaptive, meaning students of all performance levels will respond to items that are aligned to the state's content standards, but the questions that each student is offered are 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.

In the current study, the term "proficient" refers to the level of state test performance tied to federal accountability requirements, even though states may use other descriptive terms (e.g., "meets standards," "Level 3," etc.) to mean the same thing. For states that use a lower performance level for federal accountability, that lower standard is used for cross-state comparisons. Colorado, for example, has used its "partially proficient" performance level for federal accountability purposes even though it uses the higher "proficient" level of performance for internal state accountability. Similarly, New Hampshire, prior to its adoption of the New England Common Assessment Program (NECAP), used the "basic" level of performance but currently uses the "proficient" level of performance on NECAP for federal accountability reporting.

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. We use an equipercentile equating procedure, which is commonly used to compare the scales employed on achievement tests, to estimate the cut scores for 35 state instruments on a single scale. (For more information about the estimation techniques, see Appendix F.) For twenty of these states, estimates of the proficiency cut scores could be made at three points in time (generally 2002-03, 2005-06, and 2009-10), although only for certain subjects and grades. An additional six states had two points of data (generally 2005-06 and 2009-10). The remaining 11 states only had data for a single point in time. Thirteen states were not analyzed because they did not have enough students in the NWEA sample to be included.

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, 2008). 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 the MAP test, as well as our linking procedure and research methodology, is included in the Full Methodology Appendix F.

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 37 states included in the study. For some grade levels, we were not able to estimate cut scores for every state, generally because of insufficient sample size. Second, the most recent cut score estimate was also compared to a prior cut score estimate 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.

FINDINGS: The State of Proficiency

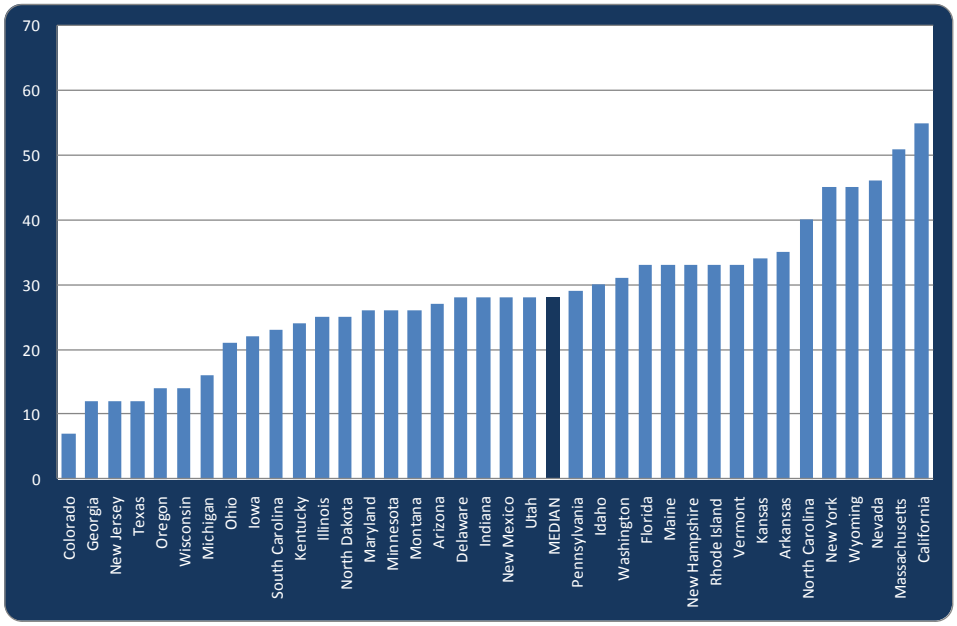
In this section we will explore five main concerns around using proficiency data to make educational decisions:

- State tests vary greatly in their difficulty, or "effective cut score"
- Within a state, difficulty varies by grade level
- Some subjects are more difficult to pass than others
- There is a clear relationship between effective cut score and proficiency rate
- States' cut scores during NCLB have varied

State tests vary greatly in their difficulty

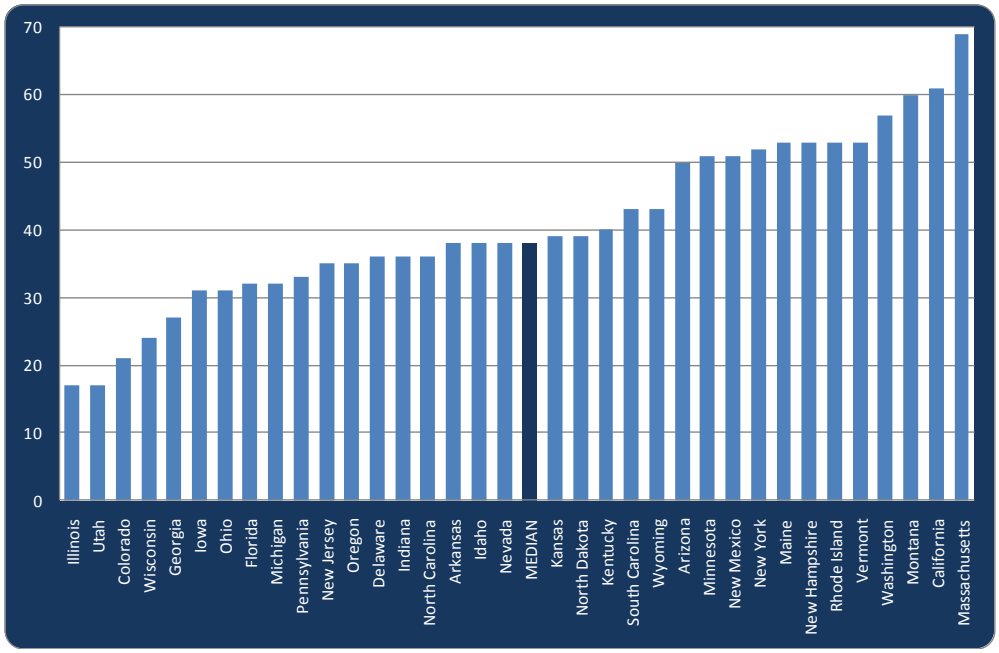
When the ESEA was reauthorized in 2001 as NCLB, there was a deliberate decision to allow each state to set its own standards and measure progress against those standards in its own way. Not surprisingly, state standards and the designated proficiency cut scores on the state tests vary significantly by state. The following figures rank states in order from easiest to most difficult for a particular subject and grade based on the estimated NWEA percentile score needed to pass the test in each state. Figure 1 depicts grade three reading proficiency cut score estimates used for federal accountability in 37 states, and Figure 2 depicts grade eight mathematics proficiency cut score estimates for 35 states. In these figures, cut scores are expressed as percentile ranks based on NWEA norms, such that higher numbers indicated more difficult proficiency standards. For example, a third grade reading standard of 7 indicates that 93% of the third graders in NWEA's normative sample of third graders would be able to pass a test of that difficulty level, whereas only 23% of the normative population would be expected to pass a test set at the 77th percentile. Cut score estimates for every state, subject, and grade are included in Appendix B.

Figure 1: Grade 3 Reading Proficiency Cut Scores for 2010 (Ranked by MAP Percentile)

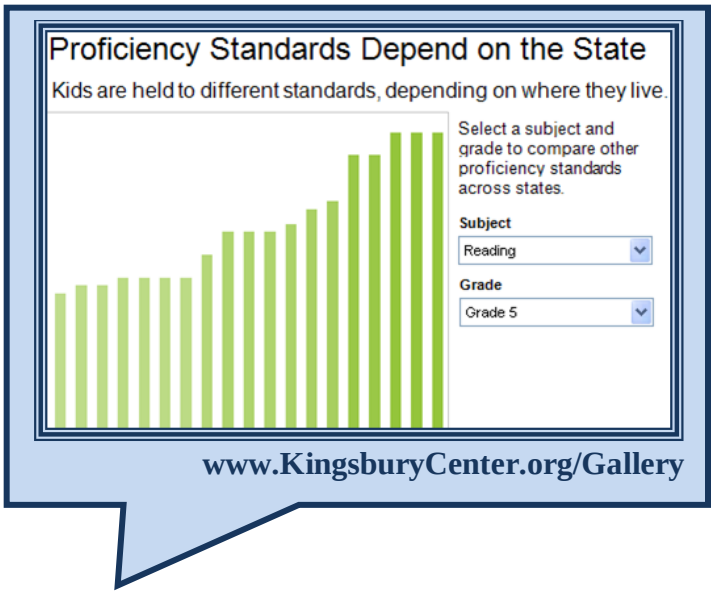


For grade three reading, the percentile estimate required to pass the state test ranged from the 7th percentile in Colorado to the 55th percentile in California. In all except two of the 37 states studied, the proficiency cut score was below the 50th MAP percentile, and in 11 states it was below the 25th percentile.

Figure 2: Grade 8 Mathematics Proficiency Cut Scores for 2010 (Ranked by MAP Percentile)



For grade eight mathematics, the percentile estimate required to pass the state test ranged from the 17th percentile in Illinois to the 69th percentile in Massachusetts. In 24 of the 35 states studied, the grade eight mathematics proficiency cut score was below the 50th MAP percentile, and in four states it was below the 25th MAP percentile.



Rankings of the state proficiency standards for each subject and grade can be seen in our online gallery.

Because state cut scores differ by grade, different patterns emerge when cut scores for each grade are aggregated to a single state average. Grade-level cut scores (percentile ranks) were converted to a normal curve equivalent score, then averaged for each state and converted back to a cut score value. The following figures show the distribution of states according to the difficulty of their proficiency cut scores across grades three through eight, with Figure 3 showing mathematics cut scores and Figure 4 showing reading cut scores.

Figure 3: Average Mathematics Proficiency Cut Scores Across All Grades (Higher = More Difficult)

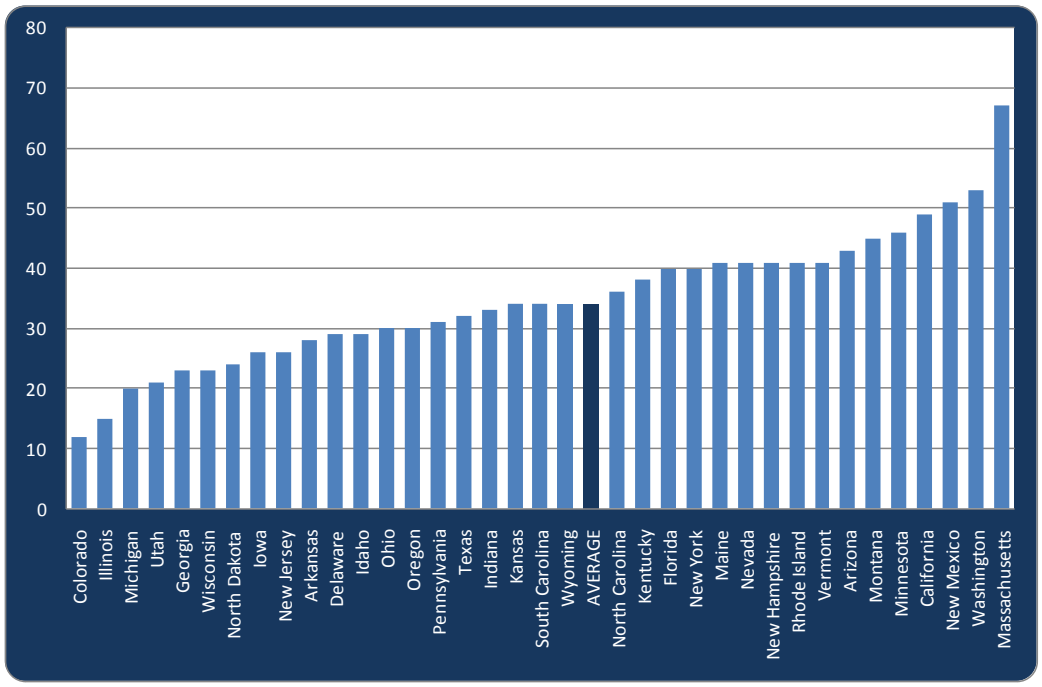
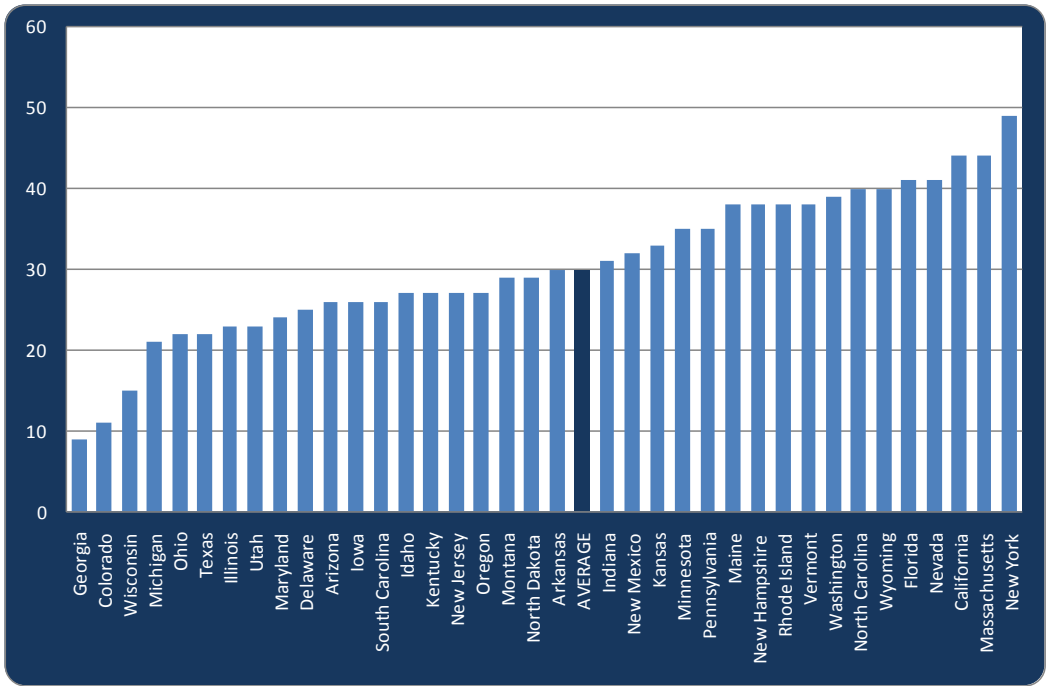


Figure 4: Average Reading Proficiency Cut Scores Across All Grades (Higher = More Difficult)



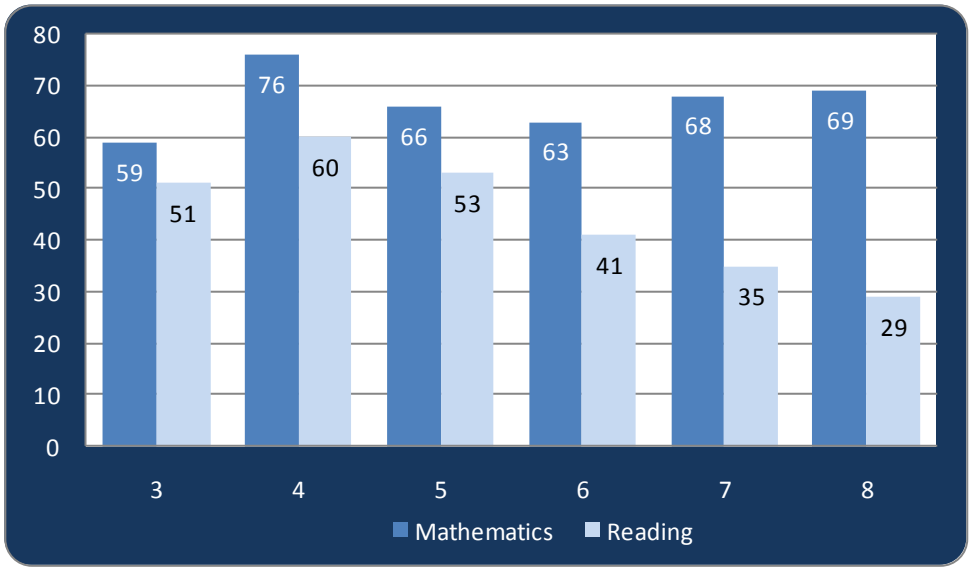
Within a State, Difficulty Varies by Grade Level

The previous section illustrated how the level of difficulty varies between states, but even within a single state there is variation in the relative difficulty across grade levels. In most states, the process for designing a test includes convening a group of teachers to develop content standards that reflect what should be learned in every subject and grade. Often, these processes happen independently of each other so that the content standards for third grade math and fourth grade math are not necessarily related.

Figure 5 shows the difference in difficulty of the Massachusetts proficiency cut scores in reading and math for grades three through eight. As can be seen, the mathematics proficiency standards range from the 59th percentile at third grade to the 76th percentile at fourth grade. Reading proficiency standards range from the 29th percentile at grade eight to the 60th percentile at grade four.

We use the term “calibration” to refer to the degree to which the proficiency standards across grades have similar relative difficulties. In Massachusetts, math proficiency standards are reasonably well calibrated, meaning that they have similar relative difficulties. Eighth grade math content may be more advanced than fifth grade content, but the typical fifth grader must put forth about the same amount of effort to be “proficient” in math as the typical sixth, seventh, or eighth grader. Reading standards in Massachusetts are not well calibrated across grades, however, with the proficiency standards at third and fourth grade more difficult than at seventh and eighth grades. Again, this does not mean that the reading content at eighth grade is less complex, but that the typical eighth grader will meet the standard much more easily than the typical third grader.

Figure 5: Proficiency Cut Score Estimates for Reading and Mathematics, 2010 (Massachusetts)

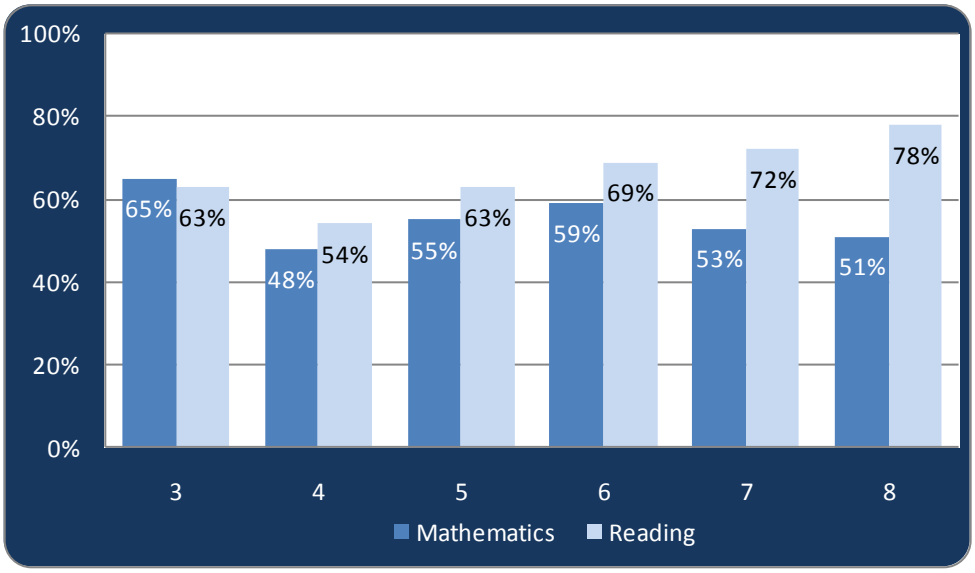


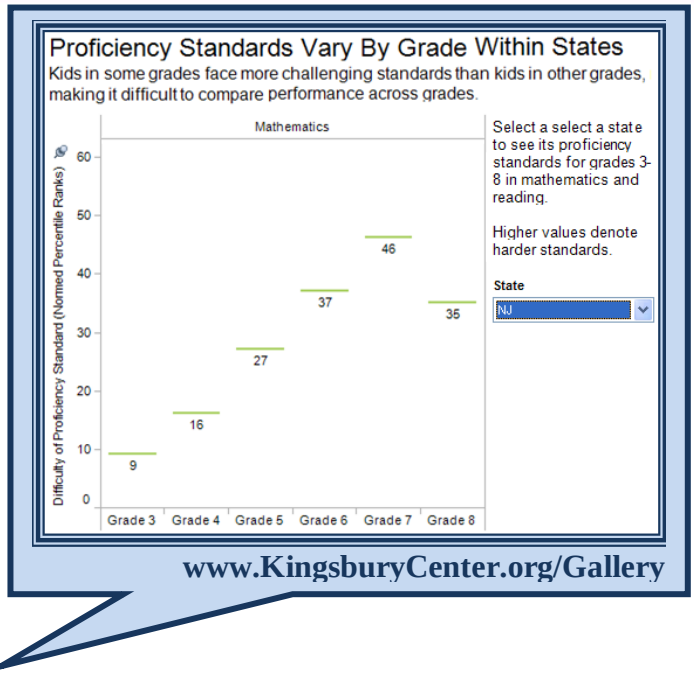
There are two primary advantages to using calibrated standards. The first advantage is that a student's performance in grade three provides an indication of her/his likely performance in subsequent years. The student who meets standards in third grade is likely “on track” to be proficient in subsequent years. A

student who fails to meet proficiency standards in third grade is "at risk" for failing later on as well. With non-calibrated standards, this is simply not the case. Meeting (or failing to meet) standards in one year conveys little information about likely future performance because the proficiency standards are entirely unrelated. A second advantage to calibrated standards is the fact that they provide more relevant information about student performance for making decisions about school improvement. When a state has a relatively hard reading proficiency standard at third grade and a relatively easy reading standard at eighth grade, as does Massachusetts, there will be a tendency to see lower rates of reading proficiency among third graders than among eighth graders. This might lead district administrators to mistakenly infer that there are problems or shortcomings at the third grade level that must be addressed in order to achieve proficiency rates that are comparable to eighth grade ones. In fact, the third graders may be performing as well or better than the eighth graders; the discrepancies in performance are merely an artifact of the lack of calibration of standards.

Figure 6 illustrates this issue, showing the percentage of students in grades three through eight who met the reading and math proficiency standards in Massachusetts in the 2009-2010 school year, as reported on the Massachusetts Department of Education website. (A list of state website sources is included in Appendix C.) As can be seen when comparing figures 5 and 6, the rates of proficiency are highest for the grades with the easiest standards and lowest in grades with the most difficult standards.

Figure 6: State-Reported Proficiency Rates in Reading and Mathematics, 2010 (Massachusetts)





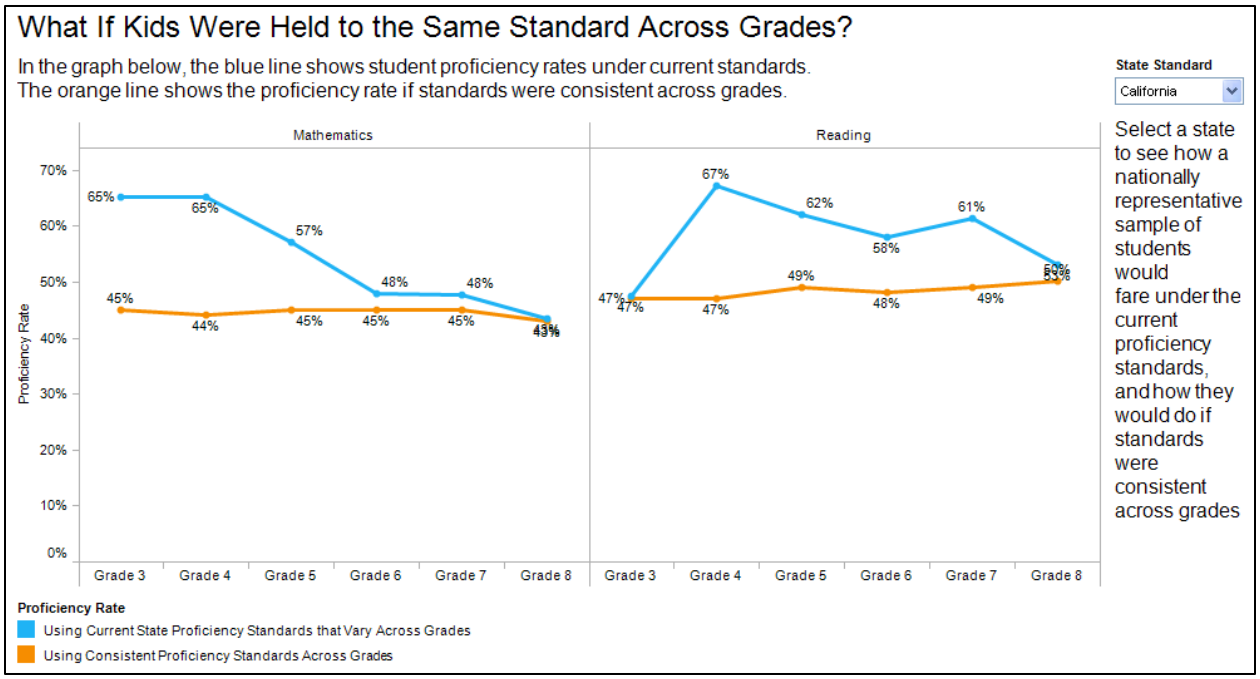
Proficiency and calibration data are shown by state, subject and grade in our online gallery.

Non-Calibrated Standards Create Challenges

One of the primary purposes of this study is to point out the challenges that arise when states have different proficiency standards across grades and subjects. For instance, imagine that a local newspaper article described a 47% passing rate in eighth grade math in comparison to the 89% passing rate in third grade reading. School administrators or parents might subsequently conclude that more resources such as curriculum and teacher training should be directed toward eighth grade math classrooms. In fact, the differences in proficiency rates are primarily a result of uneven proficiency standards. Figure 7 shows what the rates would look like in each subject and grade if a consistent standard were applied.

Among the millions of students taking NWEA tests, we extracted a sample of over 400,000 students from across the county who mirrored the percentages of the United States school-age population in free and reduced lunch percentage, racial/ethnic distribution, grade level, and urban/rural geographical location. This sample of students in grades three through eight was then evaluated against each state's proficiency standard twice: once using the cut scores set by the state, and once using the single most difficult cut score set by the state. In the figure below, which is an example taken from our online data gallery, the blue line represents the percentage of the nationally represented sample that would be considered proficient based on the state's cut scores. The orange lines show that same sample of students held to the most difficult standard across all grades.

Figure 7: Showing Consistent Standards Across All Grades and Subjects

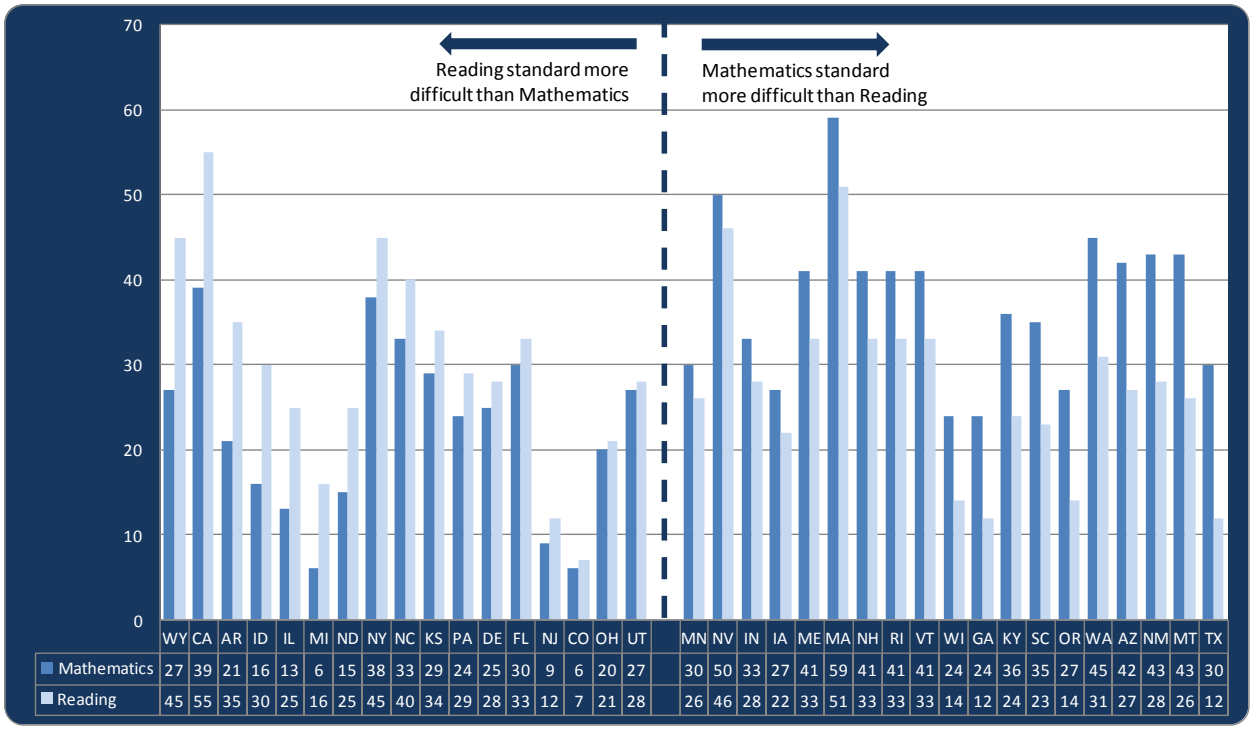


In the example above, 65% of the nationally representative sample of third grade students would be considered passing the math test using California's third grade cut score (see blue line). However, if the most difficult California cut score (eighth grade) was used, only 45% of students would be considered passing (see orange line). The public information portrayed by the blue line is what policymakers use to make decisions—in this example, California policymakers might decide that since 65% of students are passing math in the third grade and only 43% are passing in the eighth grade, they should direct more resources to the eighth grade. If those same students were evaluated using the same proficiency standard, however, then only 45% of third graders would be considered passing and 43% of eighth graders would be considered passing—not worthy of redirecting resources.

Some Subjects Are More Difficult to Pass Than Others

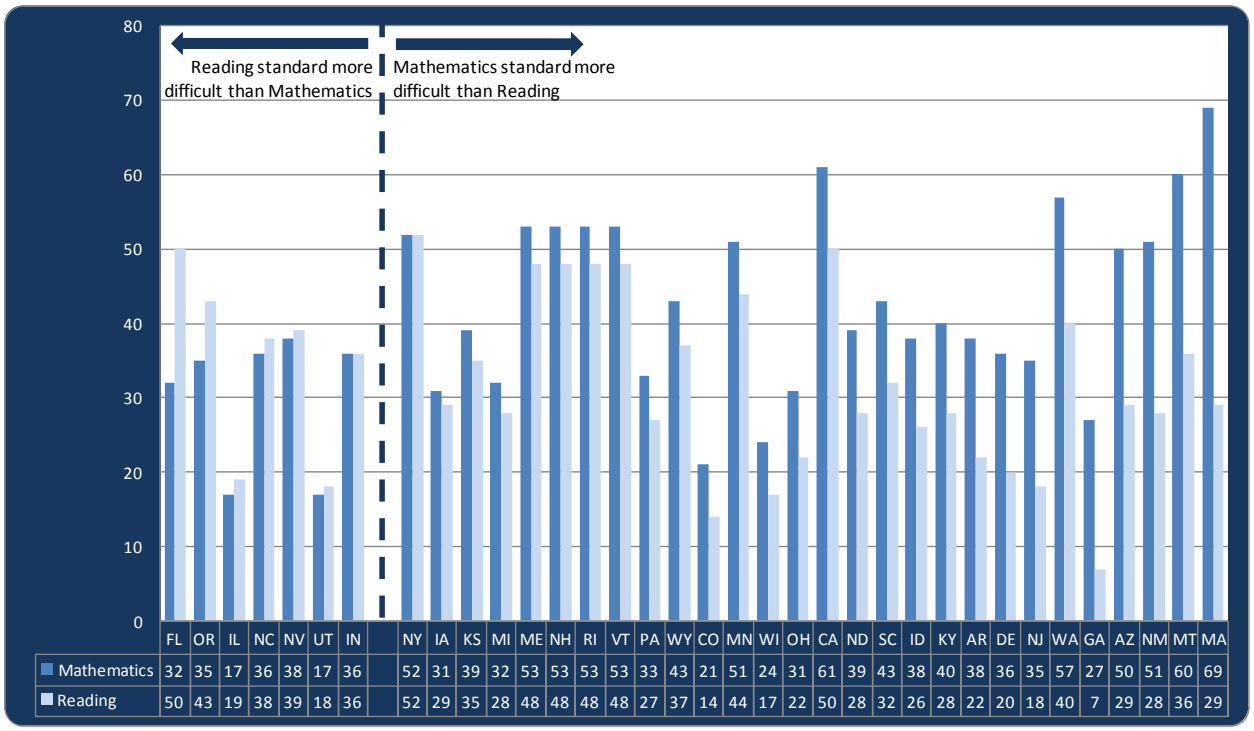
As states review their reading and math proficiency rates, they may come to the conclusion that students are doing better in one subject when compared to another, and may make decisions based on that conclusion. In most states, the percentage of students passing the state test is higher for reading than it is for mathematics, which may lead state officials to invest more resources in improving math education. The difference in proficiency rates, however, may simply be an artifact of the differences in relative difficulty for the two subjects. Figures 8 and 9 show the difference in difficulty between the states' reading and math cut scores for the third and eighth grades, respectively.

Figure 8: Grade 3 Reading and Mathematics Proficiency Estimates (Ordered by Size of Difference as Shown by MAP Percentile)



	WY	CA	AR	ID	IL	MI	ND	NY	NC	KS	PA	DE	FL	NJ	CO	OH	UT		MN	NV	IN	IA	ME	MA	NH	RI	VT	WI	GA	KY	SC	OR	WA	AZ	NM	MT	TX
Mathematics	27	39	21	16	13	6	15	38	33	29	24	25	30	9	6	20	27		30	50	33	27	41	59	41	41	41	24	24	36	35	27	45	42	43	43	30
Reading	45	55	35	30	25	16	25	45	40	34	29	28	33	12	7	21	28		26	46	28	22	33	51	33	33	33	14	12	24	23	14	31	27	28	26	12

Figure 9: Grade 8 Reading and Mathematics Proficiency Estimates (Ordered by Size of Difference as Shown by MAP Percentile)



	FL	OR	IL	NC	NV	UT	IN		NY	IA	KS	MI	ME	NH	RI	VT	PA	WY	CO	MN	WI	OH	CA	ND	SC	ID	KY	AR	DE	NJ	WA	GA	AZ	NM	MT	MA
Mathematics	32	35	17	36	38	17	36		52	31	39	32	53	53	53	53	33	43	21	51	24	31	61	39	43	38	40	38	36	35	57	27	50	51	60	69
Reading	50	43	19	38	39	18	36		52	29	35	28	48	48	48	48	27	37	14	44	17	22	50	28	32	26	28	22	20	18	40	7	29	28	36	29

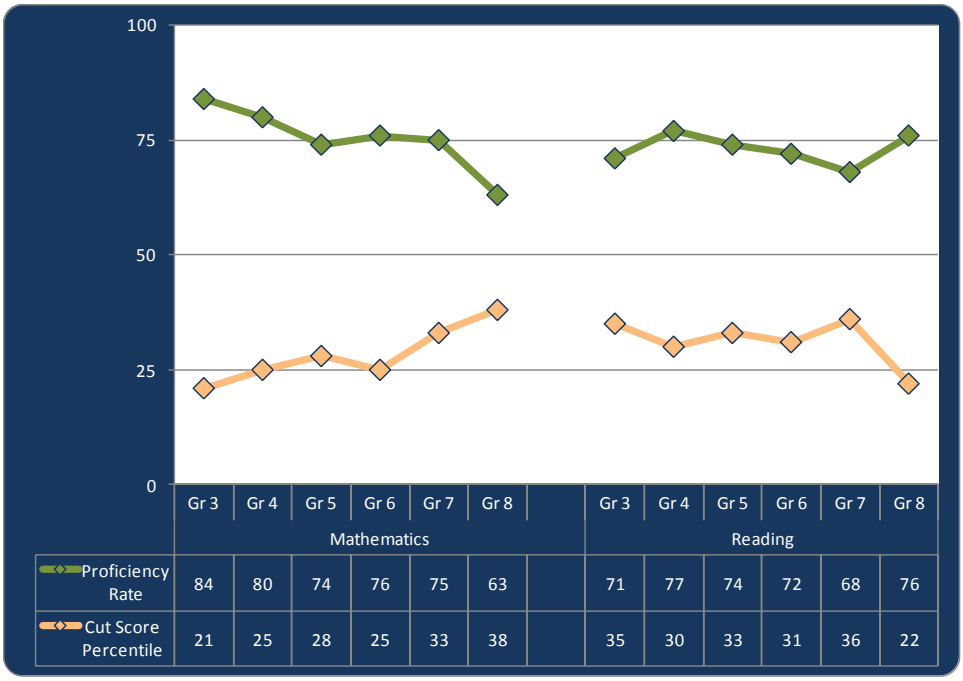
Figures 8 and 9 show that at third grade, there are roughly equal numbers of states with harder math than reading standards, but that by eighth grade, most states set much harder proficiency standards for math than for reading.

In many cases, the difference in difficulty between the reading and math standards was substantial. In Massachusetts, for example, the eighth grade math standard was set at the 69th percentile while the eighth grade reading standard was set at the 29th percentile. Put another way, one would expect that 71% of a normative sample of eighth graders would meet Massachusetts reading standards, while only 31% of that same normative sample would meet Massachusetts proficiency standards for math.

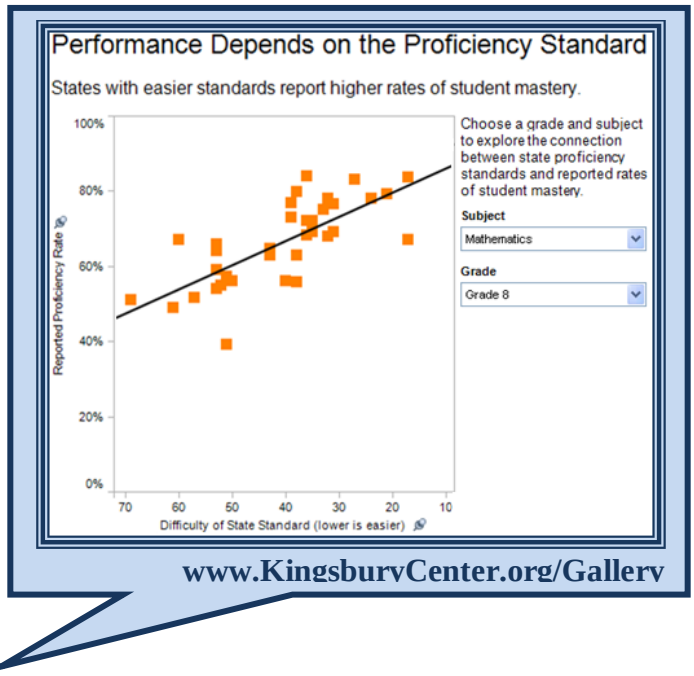
There Is a Clear Relationship Between Cut Score and Proficiency Rate

The prior three sections have all shown examples of how the proficiency cut scores set by states impact the percentage of students who pass the state test. This relationship is explored more directly in Figure 10, which plots the state-reported proficiency rates for Arkansas in reading and math directly alongside the estimated Arkansas cut scores. As can be seen, proficiency rates are higher when the proficiency standards are lower (easier). Conversely, more difficult cut scores produce lower proficiency rates. Examining this relationship across all available grades and subjects, the correlation (Pearson's r) coefficient between the estimated cut score and the proficiency rates was -.771, meaning that about 59.44% of the variation in reported proficiency rates could be explained by the difficulty of the proficiency standard. For some grades and subjects, this correlation was even higher. While policymakers and the public focus on proficiency rates, they may not realize that these rates are largely determined by the proficiency standards set by the state.

Figure 10: Proficiency Rates Vary Based on Cut Score Percentile (Arkansas)



	Mathematics							Reading					
	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8		Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8
Proficiency Rate	84	80	74	76	75	63		71	77	74	72	68	76
Cut Score Percentile	21	25	28	25	33	38		35	30	33	31	36	22



Correlations between state-reported proficiency rates and proficiency standards for each subject and grade can be seen in our online gallery.

Changes in Cut Scores During NCLB Have Varied

With the launch of the NCLB in 2001, measuring the percentage of students who are proficient on state assessments has been thrown into the national spotlight. Over the course of the decade since its introduction, states have made changes to their assessments, as well as changes to the cut scores required for students to be considered as "passing" the assessment. In *The Proficiency Illusion,* published in 2007, NWEA noted the changes in cut scores and percent proficiency for states between 2002 and 2006. Four years later, more data are available to compare changes up until the 2009-2010 school year.

Figure 11 shows the NWEA percentiles associated with each state's reading and mathematics cut scores in the most recent time period (mostly the 2009-2010 school year) and the prior time period (mostly the 2005-2006 school year). Change between the two time periods is indicated with colored arrows: a green up arrow indicates increase of six or more percentile points, a red down arrow indicates decrease of six or more; and a blue horizontal arrow indicates a change of less than six points.

Figure 11: NWEA Percentile Associated With State Standard (Higher is Harder)

MATH	2010						2006						Change 2006-2010					
	3	4	5	6	7	8	3	4	5	6	7	8	3	4	5	6	7	8
Arizona	42	38	41	46	46	50	30	28	33	40	36	42	↑	↑	↑	↑	↑	↑
California	39	39	47	57	57	61	46	55	57	62	59		↓	↓	↓	↔	↔	
Colorado	6	9	13	15	16	21	6	8	9	16	19	25	↔	↔	↔	↔	↔	↔
Idaho	16	23	31	35	36	38	50	52	48	50	47	47	↓	↓	↓	↓	↓	↓
Illinois	13	13	16	16	17	17	20	15	20	20	19	20	↓	↔	↔	↔	↔	↔
Indiana	33	36	28	31	35	36	35	32	31	27	26	34	↔	↔	↔	↔	↑	↔
Kansas	29	32	34	32	44	39	30	34	35	33	45	38	↔	↔	↔	↔	↔	↔
Maine	41	35	34	44	44	53	43	46	46	52	54	53	↔	↓	↓	↓	↓	↔
Massachusetts	59	76	66	63	68	69	68	77	70	67	70	67	↓	↔	↔	↔	↔	↔
New Jersey	9	16	27	37	46	35	13	23	26	40	43		↔	↓	↔	↔	↔	
New Mexico	43	43	51	60	61	51	46	49	54	60	61	56	↔	↓	↔	↔	↔	↔
North Dakota	15	24	21	25	28	39	20	27	23	32	39	41	↔	↔	↔	↓	↓	↔
Ohio	20	31	40	33	32	31	39	31	51	38	43	34	↓	↔	↓	↓	↓	↔
South Carolina	35	27	32	34	36	43	71	64	72	65	68	75	↓	↓	↓	↓	↓	↓
Texas	30	34	24	34	41		30	34	24	35	41		↔	↔	↔	↔	↔	↔
Washington	45	52	56	58	54	57	36	46	48	57	59	56	↑	↑	↑	↔	↔	↔
Wisconsin	24	22	22	25	23	24	29	29	26	21	21	23	↔	↓	↔	↔	↔	↔

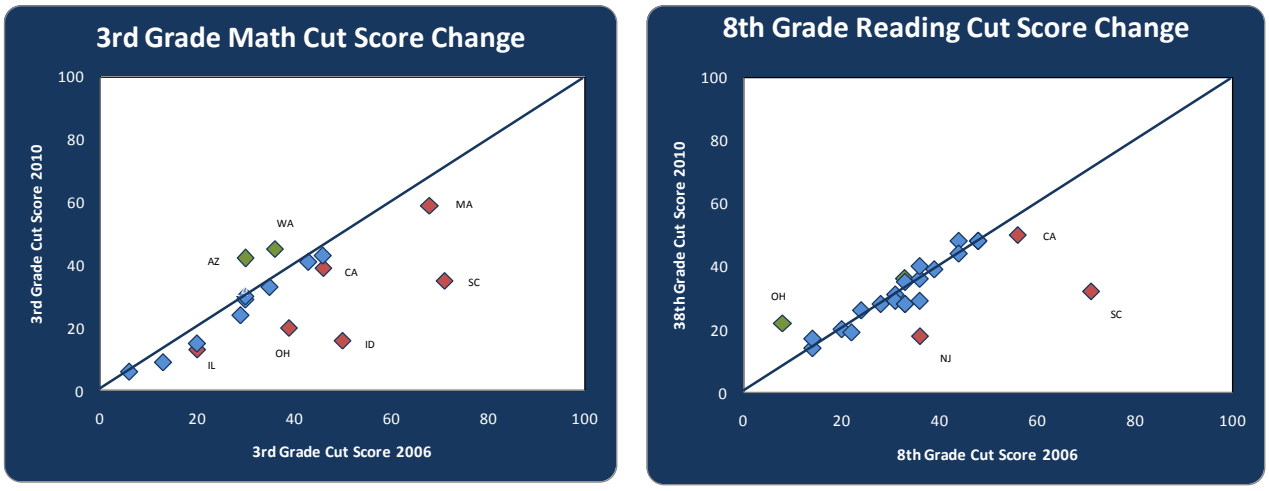
READING	2010						2006						Change 2006-2010					
	3	4	5	6	7	8	3	4	5	6	7	8	3	4	5	6	7	8
California	55	35	41	45	42	50	61	43	53	56	52	56	↓	↓	↓	↓	↓	↓
Idaho	30	27	27	27	30	26	33	29	27	27	27	24	↔	↔	↔	↔	↔	↔
Illinois	25	26	27	20	26	19	35	27	32	25	32	22	↓	↔	↔	↔	↓	↔
Kansas	34	29	40	33	33	35	35	29	40	32	32	33	↔	↔	↔	↔	↔	↔
Maine	33	34	34	43	40	48	37	43	44	46	43	44	↔	↓	↓	↔	↔	↔
Massachusetts	51	60	53	41	35	29	55	65	50	43	46	31	↔	↔	↔	↔	↓	↔
New Jersey	12	18	43	48	35	18	15	25	16	27	23	36	↔	↓	↑	↑	↑	↓
New Mexico	28	32	27	49	35	28	33	32	30	43	32	33	↔	↔	↔	↑	↔	↔
North Dakota	25	29	40	34	23	28	22	29	34	37	30	33	↔	↔	↑	↔	↓	↔
Ohio	21	21	21	25	23	22	12	11	15	7	18	8	↑	↑	↑	↑	↔	↑
South Carolina	23	26	19	30	30	32	43	58	64	62	69	71	↓	↓	↓	↓	↓	↓
Washington	31	39	37	44	47	40	37	23	27	40	49	36	↓	↑	↑	↔	↔	↔
Wisconsin	14	15	18	18	14	17	14	16	16	16	17	14	↔	↔	↔	↔	↔	↔

Green up arrow indicates increase of six or more percentile points;
Red down arrow indicates decrease of six or more;
Blue horizontal arrow indicates a change of less than six points

As is seen in Figure 11, the difficulty of the proficiency standards remained about the same for most states between the two time periods. This can also be seen In Figure 12, which shows the difficulty of the proficiency standards for third grade math and eighth grade reading in 2006 and 2010. The diagonal lines in the two figures represent a situation where proficiency standards remained at the same level of difficulty during the two time periods. The green dots above the line show states whose proficiency standards grew harder, while the red dots below the line show states whose proficiency standards grew easier in the last four years. It is interesting to note that, while the majority of proficiency standards remained at about the same level of difficulty, among those that changed, harder standards tended to get easier, and the easier standards tended to get harder.

It must be emphasized that the changes in difficulty do not always represent official cut score changes by the state departments of education, though in some cases such changes were made. These figures show the change in relative difficulty of the test, which may be the result of an official cut score change or a number of other factors. These reasons will be addressed in the following section.

Figure 12: NWEA Percentile Associated With State Proficiency Standard in 2006 and 2010



In addition to showing change between the two time periods, there are some states for which we have measurements for three time periods: typically the 2001-2002 school year in addition to the 2006-2007 and 2009-2010 school years. Cut score estimates for the earliest time period are limited, so the example below shows only a single grade.

Figure 13: NWEA Percentile Associated With Fifth Grade Proficiency Standard (Higher is Harder)

MATHEMATICS	2002	2006	2010	Direction of Change 2002-2006	Direction of Change 2006-2010
Arizona	15	33	41	↑	↑
California	49	57	47	↑	↓
Colorado	5	9	13	↔	↔
Idaho	61	48	31	↓	↓
Illinois	12	20	16	↑	↔
New Mexico	48	54	51	↑	↔
North Dakota	12	23	21	↑	↔
South Carolina	69	72	32	↔	↓
Texas	35	24	24	↓	↔

READING	2002	2006	2010	Direction of Change 2002-2006	Direction of Change 2006-2010
Arizona	13	25	29	↑	↔
California	46	53	41	↑	↓
Colorado	7	11	13	↔	↔
Idaho	22	27	27	↔	↔
Illinois	29	32	27	↔	↔
Indiana	32	29	33	↔	↔
New Mexico	30	30	27	↔	↔
North Dakota	31	34	40	↔	↑
South Carolina	52	64	19	↑	↔

Green up arrow indicates increase of six or more percentile points;
Red down arrow indicates decrease of six or more;
Blue horizontal arrow indicates a change of less than six points

As can be seen in Figure 13, the difficulty of the fifth grade proficiency standards in reading and math remained about the same (within six percentile points) for most states. Five of the nine states saw their math proficiency standards get harder between 2002 and 2006, while only one got harder between 2006 and 2010. Between 2006 and 2010, only one of nine states saw increases in the difficulty of their reading proficiency standard, while three states' standards grew easier and five remained about the same.

California's proficiency standards in reading and math grew more difficult between 2002 and 2006, but by 2010, these had decreased to lower than 2002 levels. South Carolina, which had one of the highest state standards in 2006 decreased their cut scores considerably between 2006 and 2010 when they switched from the PACT test to the PASS test; in fifth grade math, for example, the proficiency standard was at the 72nd percentile, but by 2010, that standard had decreased in difficulty to the 32nd percentile. For three of the nine states, the fifth grade reading standard stayed within six percentile points in all three time periods. For mathematics, however, only Colorado's proficiency standard remained at a roughly consistent difficulty level across the time periods.

Deliberate Changes to State Cut Scores

This study has shown that the difficulty of state proficiency standards can change over time, but so far has not addressed the *reasons* for that change. Some differences over time are due to deliberate changes made

by the state to its assessment system, while other changes over time could be due to shifting student populations, changes in student performance on the state test that do not generalize to their performance on MAP, practice effects, or other factors. While some of this information may not be known, it is important to differentiate between deliberate state policy changes to raise or lower standards and other factors which resulted in a change in estimated proficiency cut scores.

The Center on Education Policy (CEP) published a report called *State Profiles for State Test Score Trends through 2008-09* which included a list of changes that states made to their assessment systems between 2002 and 2009. A full list of the information from this report with additional information from the Michigan Department of Education website and the North Carolina Department of Public Instruction website is included in Appendix E.

CEP determined that all but one state had made substantive changes to their assessment systems between 2002 and 2009. Florida was the only exception. CEP found that the other 49 states had implemented a variety of changes to their assessment systems through this decade. Examples of changes include raising or lowering proficiency cut scores, updating the content standards on which assessments are based, implementing a new test or contracting with a new testing vendor, adding subjects and grades tested, and other changes. The three most prevalent changes CEP found in their study are summarized below:

Twelve states were identified as having changed their proficiency cut scores during this time period: Arizona in 2005, Delaware in 2006, Hawaii in 2007, Illinois in 2006, Indiana in 2009, Kansas in 2006, Michigan in 2011, Missouri in 2006, New York in 2010, North Carolina in 2008, Oregon in 2007, and Utah in 2009. The CEP report did not specify whether proficiency cut scores were raised or lowered, but comparing NWEA's estimates of proficiency cut scores over time indicates that some of these changes would have been increases while others would be decreases. This list is not exhaustive of the proficiency cut score changes, however, because several states made changes to their entire assessment systems, which resulted in modified cut scores but were not categorized by CEP as cut score changes. For instance, South Carolina had one of the highest proficiency cut scores in the nation in 2006, but an average proficiency cut score in 2010—this difference was due to implementation of the PASS assessment system in 2009.

Sixteen states were identified as having changed their content standards, but this may not accurately represent all states since the CEP report focused on changes to the assessment system as opposed to changes to content standards. For example, Oregon updates its subject area standards on a rotating annual basis and yet was not identified in the study as having made changes to its standards.

Thirty-two states were identified as having implemented a new test or new testing vendor. In some cases, these changes were accompanied by changes to the content standards and/or the proficiency cut scores. Often, when new tests or new testing systems are implemented, scores and rates cannot be compared before and after the change. In some cases, however, states made an effort to equate tests before and after changes such as these.

The No Child Left Behind Act was designed to allow states flexibility in adopting standards and assessments that best meet the needs of their state's students. Over time, it is understandable that states would want to make modifications to their assessments when it is in the best interest of education for students. The impact of these individual state decisions, however, is that trends cannot be compared over time without equating cut scores on a single scale. By doing this, NWEA has shown the *effective difference* in how states' proficiency cut scores compare to each other over time. This effective difference, however, does not always reflect an intentional policy change on the part of the state.

DISCUSSION: What the State of Proficiency means for the future

Summary of findings

When NWEA published *The Proficiency Illusion* in 2007, the No Child Left Behind legislation was operating in full force. The Bush administration that had reauthorized the Elementary and Secondary Education Act as NCLB in 2001 was in its final year, and the Obama administration had not yet been formed. Now in 2011, the educational landscape is shifting. With the 2014 deadline for all schools to be 100% proficient looming ever closer, changes to the ESEA have been proposed by the Obama administration. Such proposals include loosening the 100% accountability requirements of NCLB in order that states might focus their energies on turnarounds for the lowest performing schools, placing less emphasis on the measurement of proficiency and more emphasis on the establishment of "college and career ready" standards, and greater emphasis on measuring growth over time.

As encouraging as these proposals sound, there are additional issues related to the concept of proficiency that must be considered within the ESEA reauthorization, if we are to make progress toward ensuring that all students finish high school with the necessary skills to be competitive in the academic and professional worlds.

1. Proficiency standards vary across states, but in nearly all states studied, they remain far below any level that would be characterized as college readiness.

The fact that cut scores vary by state is not unexpected; indeed, NCLB was designed intentionally to allow this. Having different cut scores in different states means, however, that student performance cannot be compared meaningfully among states. Saying that 92% of Colorado third graders passed their state math test and 65% of California third graders passed their state math test does not mean that Colorado students are performing better than California students. One can only compare performance if the measurements can be translated into a single, comparable scale. The Common Core Standards will address part of this issue in that they are developing shared content standards for states. However, this is only the first step. Before testing data can be used effectively to make improvements to educational policy and practice, assessments to measure students' mastery of those standards must also be measured on a comparable scale.

Under NCLB, states set their own definitions for "proficiency" and this is not likely to change with ESEA's reauthorization. However, almost no states currently maintain proficiency standards that could be considered sufficient to academically prepare students for college-level studies. As shown in Figure 2, the median eighth grade math proficiency standard among all states currently examined was set at about the 38th percentile. This is not to say that states' academic content standards are low, merely that the passing scores on state tests that denote student "mastery" are, almost without exception, so low as to be meaningless in any objective sense. This will not change as states adopt the common core curriculum standards, since academic content standards are independent of proficiency standards. In order for states to establish "college and career ready" standards, there needs to be much more discussion about what this means, and whether a single standard is even desirable.

2. Standards remain uncalibrated across grades and subjects. Proficiency cut scores in the upper grades are frequently more demanding, sometimes far more demanding than cut scores in the early grades. Math cut scores are also frequently higher than reading cut scores.

The lack of an objective definition of "proficiency" or "college/career ready" means that within a state, there is no clear end point against which students can be evaluated on annual standardized tests. Without an established target, the definitions for proficiency or student mastery will not be consistent. In other words: What have students demonstrated they are proficient at doing? Such inconsistent standards are seen in nearly every state, with some grades setting excessively easy proficiency standards while other grades have harder standards. When states set inconsistent, non-calibrated standards, students' mastery in one year indicates little about their likely performance in future years, and sends confusing messages to parents, teachers, and other stakeholders about whether students are really on track for future success.

In the third grade, about half of the states had more difficult standards in reading and about half had more difficult standards in mathematics. The differences between the two subjects, however, were sometimes dramatic. For instance, in Texas the passing score for the mathematics test was 30 while the passing score for the reading test was 12. In Wyoming, the passing score for reading was 45 while the passing score for mathematics was 27. (Scores are expressed as a percentile ranking on NWEA's national percentile scale.)

In the eighth grade, the differences were even more dramatic. In 26 states, the eighth grade mathematics standard was more difficult than the reading standard, in some cases 30-40 percentage points different. In only six states was the reading standard more difficult than the mathematics standard, and only 1-18 percentage points more difficult.

There may be legitimate reasons why a subject or a grade level should be more difficult for students to pass, but there needs to be transparency about that decision. For instance, a state may want to increase the number of science, technology, engineering, and math (STEM) professionals in the workforce, and thus they may choose to have a higher state standard for mathematics than for reading. If this is a deliberate decision, however, it needs to be communicated when test results are delivered so the general public can understand the intention and the result.

3. Most of the differences in proficiency rates that are seen across states, and across grades within states, are a function of the difficulty of the state tests, not differences in student performance.

States and/or grades with easy standards have high rates of student proficiency. States and grades with hard standards have lower rates of proficiency. In general, when there are substantial differences in proficiency rates between, for example, third graders and eighth graders in a state, or between math and reading proficiency rates, those differences are largely attributable to non-calibrated standards. States must be clear about whether their scales are consistent across subjects and grades when they report findings to the public.

4. In general, the difficulty of proficiency standards within states has not changed dramatically over time. However, when it has changed, it has grown harder about as often as it has grown easier.

Most states had fairly consistent cut scores in the last four years, but this has varied somewhat by state, subject, and grade, as shown in the Findings section. For instance, South Carolina had the highest standards in the nation in 2006, but lowered its reading and math proficiency standards and is now below the median among states. Arizona's proficiency standards grew more difficult in all grades for mathematics, while Idaho's and California's math standards grew easier. Ohio's reading standards grew effectively harder, while other states showed mixed results.

It is important, however, to distinguish between deliberate changes made by the state to its assessment system and other changes such as shifting student populations, increased performance based on familiarity with the state test, or other factors. Regarding changes to state assessment systems, the Center on Education Policy (CEP) determined in 2010 that all but one state had made substantive changes to their assessment systems between 2002 and 2009. These changes included a fourth of states that raised or lowered their proficiency cut scores, more than a fourth of states that updated the content standards on which assessments are based, and a third of states that implemented a new test or contracted with a new testing vendor, among other changes.

The No Child Left Behind Act was designed to allow states flexibility in adopting standards and assessments that best meet the needs of their state's students. Over time, it is understandable that states would want to make modifications to their assessments when it is in the best interest of education for students. The impact of these individual state decisions, however, is that trends cannot be compared over time without equating cut scores on a single scale. By doing this, NWEA has shown the *effective difference* in how states' proficiency cut scores compare to each other over time. This effective difference, however, does not always reflect an intentional policy change on the part of the state.

Reasons Why These Issues Are Problematic for Educational Policy Decisions

1. Misinformation could change decisions about where limited resources should be spent.

Education officials typically need to justify every dollar spent, and often must make cuts to popular programs that don't show positive data-based outcomes. Consequently, it is essential that officials use the best possible data to make these high-stakes decisions. Imagine a state that had just piloted several new programs in math and reading, and needed to make a decision about which single one to implement throughout the state. Using state test proficiency rates to make the decision, officials might see that 62% of students are now passing the reading test while only 47% of students are passing the mathematics test. Given this information, the state might choose to invest in the reading program and cut the mathematics program. Without information about the relative difficulty of the two standards, however, that decision might not produce the best outcome for student performance.

If low proficiency rates cause a state to invest in a certain program when the lower proficiency rate is just an artifact from where the cut score is set, those resources could have been used more efficiently in another grade or subject.

2. Students and parents want a clear college and career trajectory for K-12.

If the goal of public education is to prepare K-12 students for college and careers, it is imperative that K-12 standards address the entire learning trajectory. Academic content standards at each grade level must be set, not independently, but as part of a longitudinal trajectory established to ensure that students graduate in the 12th grade with appropriate knowledge and skills to attend post-secondary programs or enter careers. Student mastery at early grades should be indicative that the student is "on track" to graduate; similarly, failure in the early grades should be an indicator of "at risk" status for later years. Under the current system of non-calibrated proficiency standards, performance in early grades provides little or no information about student progress toward any objective outcome later on. The education system should be set up so students, parents, and teachers know the child's entire learning trajectory so progress can be measured against the student's long-term goals.

Recommendations for ESEA

Based on the findings from this study, as well as our original *Proficiency Illusion* study and other research we have contributed over the years, we have the following recommendations for policymakers as they reauthorize the Elementary and Secondary Education Act (known as ESEA and NCLB).

- New standards and assessment systems should be structured based on what students should know and be able to do at the end of high school, and the proficiency/mastery standards at each grade scaled accordingly so students know where they are in meeting that target. The current system of siloed, non-calibrated performance standards at every grade sets students up for failure because the learning continuum does not continue from kindergarten to college. Student, parent, and community expectations are based on college readiness; standards should be as well. Student mastery (or non-mastery) at every grade should provide information about whether students are on track for future success.

- Similarly, it should not be more difficult for a typical fourth grade student to pass the mathematics test than the reading test unless that is a clear, desired outcome. We recommend that assessment systems should be scaled and calibrated to reflect equivalent levels of difficulty across subjects or that the intention of scaling subjects differently is clearly articulated and understood.

- We recommend that the two assessment systems developed by PARCC and SMARTER Balance either use an equivalent measurement scale or that a reliable crosswalk between the two systems is readily available so students from both systems can be compared, both across states and across grades.

How we hope the data galleries will help

This report is only one part of the work we have done to portray the differences in state standards and how these differences lead to misinterpretations of student performance and misallocations of resources. We have also developed online data galleries that allow users to interact with real data to see for themselves the effects of different types of policies. These data galleries can be accessed at the following site:

http://www.KingsburyCenter.org/Gallery

We encourage all stakeholders in the educational community to check out the Kingsbury Center Data Gallery and see for themselves how data can affect policy and practice. Each data exhibit includes video clips, interactive data visualizations using real data from the study, and links to other studies and blog posts that are related to the study topic. There is also space for visitors to leave comments, ask questions, and share.

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, & 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., & Urbina, S. (1997). *Psychological Testing* (7th ed.). New York, NY: MacMillan.

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

Cronin, J., Kingsbury, G. G., Dahlin, M., Adkins, D., & Bowe B. (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.

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

Ingebo, G. (1997). *Probability in the Measure of Achievement.* Chicago, IL: Mesa Press.

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.

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). *The Impact of the No Child Left Behind Act on Student Achievement and Growth: 2005 Edition.* Lake Oswego, OR: Northwest Evaluation Association.

Northwest Evaluation Association (2008). *Technical Manual for the NWEA Measures of Academic Progress and Achievement Level Tests.* Portland, OR: Northwest Evaluation Association.

Northwest Evaluation Association. (2008). *RIT Scale Norms.* Portland, OR: Northwest Evaluation Association.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement 14* (2), 97-116.

The State of Proficiency

The Kingsbury Center at NWEA
July 2011

APPENDICES

Appendix A: Mathematics Cut Score Estimates for 2009-2010

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Arizona	42	38	41	46	46	50
Arkansas	21	25	28	25	33	38
California	39	39	47	57	57	61
Colorado	6	9	13	15	16	21
Delaware	25	26	24	29	36	36
Florida	30	40	46	52	43	32
Georgia	24	29	25	21	16	27
Idaho	16	23	31	35	36	38
Illinois	13	13	16	16	17	17
Indiana	33	36	28	31	35	36
Iowa	27	26	20	29	27	31
Kansas	29	32	34	32	44	39
Kentucky	36	34	38	40	41	40
Maine	41	35	34	44	44	53
Massachusetts	59	76	66	63	68	69
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	9	16	27	37	46	35
New Mexico	43	43	51	60	61	51
New York	38	38	37	41	38	52
North Carolina	33	33	42	35	43	36
North Dakota	15	24	21	25	28	39
Ohio	20	31	40	33	32	31
Oregon	27	27	30	37	28	35
Pennsylvania	24	27	39	34	34	33
Rhode Island	41	35	34	44	44	53
South Carolina	35	27	32	34	36	43
Texas	30	34	24	34	41	n/a
Utah	27	26	28	25	12	17
Vermont	41	35	34	44	44	53
Washington	45	52	56	58	54	57
Wisconsin	24	22	22	25	23	24
Wyoming	27	35	37	30	36	43

Appendix B: Reading Cut Score Estimates for 2009-2010

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Arizona	27	29	29	24	24	29
Arkansas	35	30	33	31	36	22
California	55	35	41	45	42	50
Colorado	7	12	13	10	13	14
Delaware	28	32	23	27	23	20
Florida	33	40	53	34	37	50
Georgia	12	12	12	7	11	7
Idaho	30	27	27	27	30	26
Illinois	25	26	27	20	26	19
Indiana	28	30	33	31	34	36
Iowa	22	22	21	35	30	29
Kansas	34	29	40	33	33	35
Kentucky	24	27	27	27	30	28
Maine	33	34	34	43	40	48
Maryland	26	20	23	23	27	31
Massachusetts	51	60	53	41	35	29
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	12	18	43	48	35	18
New Mexico	28	32	27	49	35	28
New York	45	45	50	51	53	52
North Carolina	40	37	44	39	48	38
North Dakota	25	29	40	34	23	28
Ohio	21	21	21	25	23	22
Oregon	14	18	32	33	30	43
Pennsylvania	29	34	50	36	38	27
Rhode Island	33	34	34	43	40	48
South Carolina	23	26	19	30	30	32
Texas	12	23	30	21	32	28
Utah	21	25	27	25	18	
Vermont	33	34	34	43	40	48
Washington	31	39	37	44	47	40
Wisconsin	14	15	18	18	14	17
Wyoming	45	34	40	38	50	37

Appendix C: Website References

1. **The original *Proficiency Illusion* report from 2007:**

   Executive summary:
   http://www.kingsburycenter.org/sites/default/files/Proficiency_Exec.pdf

   Full report:
   http://www.edexcellence.net/publications-issues/publications/theproficiencyillusion.html

2. **The data galleries sites:**

   Kingsbury Center Data Galleries Home: http://kingsburycenter.org/gallery

   The Proficiency Illusion Data Gallery: http://kingsburycenter.org/gallery/gallery-detail-1

   The Achievement Gap Data Gallery: http://kingsburycenter.org/gallery/gallery-detail-4

3. **Information about Common Core State Standards:**

   http://www.corestandards.org/

4. **Information about the two assessment consortia:**

   SMARTER Balanced Assessment Consortium (SBAC):
   http://www.k12.wa.us/SMARTER/default.aspx

   Partnership for the Assessment of the Readiness for College and Careers (PARCC):
   http://www.parcconline.org/

5. **Elementary and Secondary Education Act (ESEA) reauthorization:**

   ESEA Reauthorization Blueprint from March 2010:
   http://www2.ed.gov/policy/elsec/leg/blueprint/index.html

   U.S. Department of Education's blog on ESEA reauthorization:
   http://www.ed.gov/blog/topic/esea-reauthorization/

Appendix D: Sources for 2009-2010 State Proficiency Rates

- AR: http://normessasweb.uark.edu/schoolperformance/State/SRCy3.php
- AZ: http://www.ade.az.gov/srcs/statereportcards/StateReportCard2010.pdf
- CA: http://star.cde.ca.gov/star2010/ViewReport.asp?ps=true&lstTestYear=2010&lstTestType=C&lstCounty=&lstDistrict=&lstSchool=&lstGroup=1&lstSubGroup=1
- CO: http://www.cde.state.co.us/cdeassess/documents/csap/csap_summary.html
- DE: http://dstp.doe.k12.de.us/DSTPmart9/default.aspx
- FL: http://fcat.fldoe.org/fcinfopg.asp
- GA: http://www.doe.k12.ga.us/ReportingFW.aspx?PageReq=102&StateId=ALL&T=1
- ID: http://www.sde.idaho.gov/site/assessment/ISAT/results.htm
- IN: http://www.doe.in.gov/assessment/2010/
- KS: http://online.ksde.org/rcard/summary/state.pdf
- KY: http://applications.education.ky.gov/ktr/default.aspx
- MA: http://www.doe.mass.edu/mcas/results.html?yr=2010
- MD: http://www.mdreportcard.org/Assessments.aspx?WDATA=State&K=99AAAA
- ME: http://www.maine.gov/education/necap/1011necapscores/statewide.pdf
- MI: http://www.michigan.gov/documents/mde/Fall_2010_STATEWIDE_MEAP_RESULTS_349215_7.pdf
- MN: http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/index.html
- MT: http://opi.mt.gov/Reports&Data/nclb-reports.php
- NC: http://accrpt.ncpublicschools.org/app/2010/disag/
- ND: http://www.dpi.state.nd.us/dpi/reports/Profile/0910/99999.htm
- NH: http://reporting.measuredprogress.org/NHProfile/reports.aspx?view=32
- NM: http://www.ped.state.nm.us/AssessmentAccountability/AcademicGrowth/NMSBA.html
- NJ: http://www.state.nj.us/education/schools/achievement/2011/
- NV: http://www.nevadareportcard.com/
- NY: https://www.nystart.gov/publicweb-external/2010statewideAOR.pdf
- OH: http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=222&ContentID=15606&Content=95696
- OR: http://www.ode.state.or.us/search/page/?id=1821
- PA: http://www.portal.state.pa.us/portal/server.pt/community/school_assessments/7442
- RI: http://reporting.measuredprogress.org/NECAPpublicRI/select.aspx
- SC: http://ed.sc.gov/topics/assessment/scores/pass/2010/statescoresgrade.cfm
- TX: http://ritter.tea.state.tx.us/perfreport/aeis/2010/state.html
- UT: http://schools.utah.gov/assessment/Reports/Results_CRT_2010-pdf.aspx
- VT: http://education.vermont.gov/new/pdfdoc/dept/press_releases/EDU-NECAP_PowerPoint_Presentation_2010_2011.pdf
- WA: http://reportcard.ospi.k12.wa.us/summary.aspx?year=2009-10
- WI: https://apps2.dpi.wi.gov/sdpr/spr.action
- WY: https://fusion.edu.wyoming.gov/MySites/Data_Reporting/data_reporting_assessment_reports_results_state_level.aspx

Appendix E: Deliberate changes to state cut scores

Sources: The Center on Education Policy *State Profiles for State Test Score Trends through 2008-09*; the Michigan Department of Education website; and the North Carolina Department of Public Instruction website.

Alabama	2004-05: ARMT assessments implemented, replacing the Stanford-10 in grades 3, 5, and 7 in reading and in grades 3, 5, 7, and 8 in mathematics 2005-06: Displaced Hurricane Katrina students disaggregated for this administration only
Arizona	2005: Cut points reset
Arkansas	2005: Reset standards for grades 3-8 Benchmark Exams and developed a vertical scale (scales for EOC and grade 11 Literacy Exams remained unchanged)
Alaska	2005: Switched from using the Alaska Benchmark Exams (ABE) to the ASBA and expanded testing to all of the grades 3-9 2006: Switched to ASBA in grade 10
California	2003: CST revised to target only CA content standards 2006: New science tests added for grades 8 and 10 2008: CSTs expanded to include testing in ELA in grades 2-11; math in grades 2-9; science in grades 5, 8, and 10; and history/social science in grades 8 and 11
Connecticut	2005-06: Added grades 3, 5, and 7 2006: Introduced new generation of CMT, switched to spring testing 2007: Introduced new generation of CAPT
Colorado	2004: Changed from reporting AYP by grade span to reporting by specific grades 2004: Introduced math assessments in grades 3-4 but scores not used for AYP until 2005 2006: Included grades 5 and 10 in state science assessment
Delaware	Spring 2006: Cut scores for reading and math proficiency levels changed
District of Columbia	Spring 2006: Changed test to DC CAS Spring 2008: Science tests administered
Florida	None
Georgia	2008: The Georgia Performance Standards (GPS) were phased in to replace Georgia's Quality Core Curriculum (QCC); scores are changing accordingly as tests are phased in. 2008: Students in grade 11 who took the GHSGT in English language arts for the first time took a new version of the test based solely on the GPS. The GHSGT math test was still based on the QCC. 2008: New tests administered in math in grades 3-5 and 8.
Hawaii	March 2007: New HCPS III state assessment proficiency levels set and approved
Idaho	2006: Switched test vendors; new vendor designed an adaptive version of the ISAT. 2007: Standard scores were set and will remain until substantive changes are made to the standards, which will require changes to the test.

Illinois	2005-06: Changed test vendors. 2005-06: Switched to a vertical scale for scoring the test; cut scores were changed accordingly (grade 8 math scores in particular were changed after a bridge study found that cut scores were too high). 2006-07: Added another test vendor. 2008: Scoring of the PSAE was modified such that all items contribute equally to the overall score. A process was used to equate 2007 results from the old methodology to the new methodology.
Indiana	2002: Grades 3, 6, and 8 ISTEP+ tests modified to reflect new Indiana standards; vertical scale developed; cut scores/performance level descriptors introduced. 2004: New tests administered in grades 4, 5, 7, and 9. 2004: Grade 10 GQE revised to reflect new standards; first year of full administration of ISTEP+ to grades 3-10. 2008-09: For this year only, students took the ISTEP+ twice as the test moved from a fall testing window to a spring testing window. IN also made the transition to end-of-course tests in Algebra I and English 10. 2009: First spring administration of new ISTEP+. Administered in two sessions (open-ended portion in March, multiple choice in April-May). New cut scores established. Class of 2011 will be last group of students to take the current GQE. Class of 2012 will take end-of-course assessments in Algebra 1 and English 10. Cut scores will be established summer of 2010.
Iowa	2005-06: Began assessing all students in grades 3-8, 11 for inclusion in AYP reporting
Kansas	2004: State revised standards 2005-06: State expanded reading assessment to grade 2 (local choice of instrument), grades 3-8, and high school grades 9, 10, or 11 (at end of opportunity-to-learn, district-level decision); expanded math assessment to grades 3-8 and one grade in high school 2005-06: Kansas Assessment with Multiple Measures (KAMM) replaced the Kansas Assessment Program or what was known as the modified assessment 2006: State developed new cut scores and AYP targets Spring 2007: State implemented flexible "opportunity-to-learn (OTL)" testing procedures for high school reading and math; schools have the flexibility to schedule these tests after students have had an opportunity to learn the content being tested
Kentucky	2007: Changed test vendor and assessment scale
Louisiana	2005-06: iLEAP implemented to assess students in grades 3, 5, 6, 7, and 9 (replacing Iowa Tests of Basic Skills)
Maine	2005-06: Began basing assessments on revised standards; made online testing available. 2005-06: Replaced high school assessment with the SAT. 2006-07: Augmented the SAT mathematics test with state-specific items. 2006-07: Rescaled the MHSA tests in both reading and math to use an 80-point scale. 2009-10: Replaced MEA with the New England Common Assessment Program (NECAP) in a consortium with New Hampshire, Rhode Island, and Vermont.
Maryland	2005: English 2 HSA exam replaced reading 10 exam. 2008: Changed policy for reporting scores from high school exams. Instead of reporting only those scores from the first time students took the test, the state began reporting the highest scores of students who took the high school exams multiple times. 2007-2008: Maryland includes the proficient scores from the modified assessments in calculating AYP and cap the scores at 2% of the total tested population. The modified assessments are based on modified achievement standards aligned with the state's content standards.

Massachusetts	2002: New scaling system adopted. 2005-06: Reading/ELA and math tested in all of the grades 3-8 and 10. Prior to 2005-06, reading/ELA was tested in grades 3, 4, 7, and 10, and math was tested in grades 4, 6, 8, and 10.
Michigan	2002-03: Proficiency levels changed. Fall 2005: All students in grades 3-8 assessed for the first time (prior assessment included one administration in elementary school and one in middle school). 2005-06: Separate scale implemented for each grade, although standards are vertically articulated; comparisons cannot be made across grades. 2005-06: MEAP content standards revised, new standards set, and assessment window shifted from winter to fall; cannot compare these scores with scores from previous years. 2006-07: MME replaced previous high school test. 2011: February approval by MDE to raise cut scores.
Minnesota	2006: Spring test administration became baseline for equating results from reading and math MCA-II tests for all grades; new standard setting conducted in summer 2006
Mississippi	July 2001: SATP cut scores set for English II. November 2002: SATP cut scores set for Algebra I. November 2004: SATP cut scores set for Biology I and U.S. History. 2006–07: First year that MCT and SATP only were administered and previous tests were totally phased out (including Functional Literacy Exam, grades 4 and 7; Writing Assessments, and TerraNova Norm-Referenced Tests). 2006: Language Arts frameworks revised. 2007: Math frameworks revised. 2007-08: MCT2 first administered to grades 3-8; SATP2 first administered in Algebra I and English II. New cut scores set.
Missouri	2005-06: Missouri began testing all the grades from 3-8 and high school. The state also changed assessments, changed the number of achievement levels from five to four, and changed the cut scores defining proficient performance. 2008-09: End-of-course exams for course content replaced high school grade-span tests for Math grade 10 and Communication Arts grade 11.
Montana	2004: Changed from using a norm-referenced test (Iowa Test of Basic Skills) for NCLB purposes to administering criterion-referenced MontCAS tests in spring 2004 2004: Began testing grade 10 instead of grade 11 2006: Added grades 3, 5, 6, and 7 to testing
Nebraska	2008-09: Statewide reading test piloted; to be implemented in 2009-2010, with math one year later.
Nevada	2004: New test contractor chosen
New Hampshire	2005-06: New assessment system (NECAP) administered at grades 3-8 Fall 2007: New NECAP assessment administered in grade 11
New Jersey	2001: Standards set for the ESPA and NJ ASK 4 in language arts. March 2004: NJ ASK 4 replaced ESPA for accountability purposes (name changed but test content and structure remained the same). March 2005: NJ ASK 3 first used for accountability purposes. 2005-06: Grades 5, 6, and 7 added to testing. Spring 2007: HSPA science assessments began. 2008: New NJ ASK grade 5-8 programs were implemented, new standards were set. 2009: New grade 3-4 testing programs established in 2009, with standards set in July 2009.
New Mexico	Spring 2005: New tests administered in grades 3-9 and grade 11. These changes required new

	standard setting at all grade levels. For this reason, the state has been careful to not make direct comparisons between 2004 and 2005. New test was used for NCLB. In addition, grades 3, 5, 6, and 7 were tested for the first time. 2007: Changed test vendor and set new standards for high school test.
New York	2006: Students in grades 3-8 were assessed in ELA and mathematics. Prior to that, grades 4 and 8 were assessed, but NYSED advised that 2006 tests were not comparable to previous years. 2010: Cut score changes.
North Carolina	2002-03: Modified EOG reading score scale. 2005-06: Administered new EOG math assessments; in math, set new annual measurable objectives, aligned to new standards, for AYP purposes under NCLB. 2007-08: Administered new test editions for EOG Reading (grades 3-8). Established new cut scores and set new baseline for annual measurable objectives to align to more rigorous standards. 2008-09: Began using the higher of the original or retest scores for calculating state ABCs Performance Composite and AYP results for Reading Comprehension and Math in grades 3-8 and Science in grades 5 and 8. The same policy will apply to high school tests beginning in 2009-10. Data included in this profile exclude retests. *Prior to 2009, data for overall percentages proficient and above came from NC's website, while data broken down by achievement levels were provided by NC from another source. Due to different rules for suppressing small cells, and other factors, discrepancies exist. Specifically, the sum of the discrete percentages of students at Level 3 (proficient) and Level 4 (advanced) differs slightly from the percentage of students performing at or above Level 3 reported for NCLB purposes*.
North Dakota	Spring 2005: New standards set, new cut scores established
Ohio	2004: Ohio Achievement Tests implemented as replacement for state Proficiency Tests by 2006 2004: OAT cut scores established in reading and mathematics 2005: OAT cut scores established in science, social studies, and writing Spring 2005: Final administration of Proficiency Tests
Oklahoma	2009: Performance standards raised to align closer with NAEP for Grades 3-8 in Math and Reading
Oregon	2006-07: Cut scores changed for all previously tested grades, so data for 2006–07 and beyond are not comparable to those from previous years
Pennsylvania	2006-07: Revised assessment anchors based on Achieve, Inc., alignment study; formed the blueprint/test specifications for the 2007 PSSA 2008: Brought in new test contractor; conducted validation study of cut scores for grade 3
Rhode Island	2005-06: Implemented NECAP, a new assessment system developed in collaboration with Vermont and New Hampshire, in grades 3-8 (Maine joined in fall 2009); replaced New Standards Reference Exam (NSRE) tests at elementary and middle school levels.
South Carolina	2009: New testing system, PASS, implemented for grades 3-8
South Dakota	2004-05: Developed new reading and math standards and new reading assessment 2005-06: Developed new math assessment 2008-09: New reading standards and assessment developed
Tennessee	2004-05: The TCAP became strictly criterion-referenced (concordance study completed to ensure comparability with 2003-04 data)

Texas	2002-05: State phased in higher passing standards for TAKS grades 3-11. In 2003, the passing standard was two standard errors of measurement (SEM) below the panel-recommended standard; in 2004, it was 1 SEM below the panel-recommended standard; and in 2005, it was fully phased in. 2008: SDAA II, LDAA, and RPTE no longer administered; implemented TAKS-Modified, TAKS-Alternate, and TELPAS.
Utah	Spring 2003: Four new performance levels established (minimal, partial, sufficient, and substantial), replacing prior levels of mastery and non-mastery 2003-04: Standards reset for all assessments 2007: First administration of UALPA for English language learners 2009: New standards and cut scores implemented for math *Utah state education department staff identified pre-algebra for middle school and geometry for high school as the most appropriate CRT end-of-course exams to use to represent math achievement.*
Vermont	2005-06: Switched to new assessment system (NECAP), a collaboration with New Hampshire, Maine, and Rhode Island; replaced NSRE assessments
Virginia	2005-06: Grades 4, 6, and 7 were tested in reading and math and included in AYP determinations for first time 2005-06: Tests for grades 3, 5, and 8, and high school end-of-course tests were revised; data not comparable to previous years
Washington	2005-06: Testing expanded to include grades 1 and 3-8
West Virginia	2003-04: Switched to WESTEST assessment from Stanford Achievement Test-9th Edition (SAT-9). 2008-2009: Administered new WESTEST 2 assessment, which is aligned to recently adopted content standards and replaces the original WESTEST.
Wisconsin	2002-03: Test window changed to November from February. Fall 2005: Switched to WKCE-CRT (from a state-augmented version of the off-the-shelf TerraNova test); grades 3-8 and 10 assessed (previously, only grades 4, 8, and 10 were assessed). Fall 2005: Scale scores rescaled to reflect move to completely customized tests in reading and math. Proficiency standards were equated and can be compared across assessments.
Wyoming	2006: First operational PAWS assessment in grades 3-8 and 11 (formerly 4, 8, and 11 were assessed under WyCAS)

Appendix F: Full Methodology

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 kindergarten through high school that are taken by students in about 4,610 school systems in all 50 states, as well as in over 100 countries internationally.

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, & National Council on Measurement in Education, 1999). The *Guidelines for Computer-Based Testing* (Association of Test Publishers, 2000) 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 all 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 & Urbina, 1997). A test has content validity built into it by careful selection of which items to include (Anatasi & 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 5,200 items in reading and 8,000 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 2,000 items corresponding to the content standards of each state is selected.

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 40-55 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, Adkins & Bowe, 2007; NWEA, 2005b). Of the 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 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 14 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. Across the reading studies, Pearson correlations between MAP and the 14 state assessments averaged 0.79; the average correlation across the mathematics studies was 0.83. 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 20 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 measurement scales that may cause drift in a state test's difficulty over time.

Sample

We computed proficiency cut score estimates for every state with a sufficient number of students using the NWEA assessment to provide a valid match to state assessments. 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 procedure produced populations of students who had taken both their 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 most states were estimated using this method.

For a small number of tests with insufficient individual student data, an alternate procedure (method 2) was used. 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% 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% of the count taking the respective state test.

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 26 states. Of these, 19 produced cut scores for multiple test years, allowing us to examine changes over time. 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% 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% 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% of the group minus the 75% 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. *The Proficiency Illusion* found that in mathematics, compiled MAP proficiency estimates over-predicted 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% 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, 2008).

Estimates Part 2: Changes in Cut Scores Over Time

Multiple estimates were generated for 20 states, permitting comparisons of cut scores over time. The most recent estimate was taken from data gathered during the fall 2009 and spring 2010 testing terms. The prior estimates used estimates from the spring 2005, fall 2005, spring 2006, fall 2006, or spring 2007 testing terms. The first 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.

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, relatively speaking, 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 relatively easy and lower when the standard is more 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.

For the *State of Proficiency* study, we created a nationally representative random sample from a large database of student academic achievement, and used that sample to estimate the impact of calibration on observed proficiency rates across grades. Using the NCES Common Core of Data (CCD) from 2008, we totaled the number of students in each school category defined by three elements: geography (City, Suburban, Town, Rural), level (Primary or Middle), and poverty status (<25% FRL Rich, 25-50% FRL MedRich, 50-75% FRL MedPoor, >75% FRL Poor). An example of a resulting school category is CityMiddleMedPoor. Within each of these 32 school categories, the total number of students in each of five racial/ethnic groups (Asian, Black, Hispanic, Native, White) was determined for the nation. An example of a resulting student category is CityMiddleMedPoorAsian. For each of these 160 categories, a corresponding random sample of students who took a reading and math test in grades three through eight in spring 2010 was selected from NWEA's growth research database. The result was a sample of over 400,000 students that represents the same school and student characteristics of the nation. The number of students in each category appears in the table on the next page.

Level	FRL	Race	City	Suburb	Town	Rural
Middle	MedPoor	Asian	639	424	91	97
Middle	MedPoor	Black	2821	2473	1007	1258
Middle	MedPoor	Hispanic	3386	3082	1386	1153
Middle	MedPoor	Other	286	243	242	236
Middle	MedPoor	White	3256	3178	4192	4941
Middle	MedRich	Asian	633	838	176	241
Middle	MedRich	Black	1377	2280	416	1055
Middle	MedRich	Hispanic	1597	2353	723	1146
Middle	MedRich	Other	258	414	216	259
Middle	MedRich	White	4710	9052	6770	8996
Middle	Poor	Asian	611	165	23	19
Middle	Poor	Black	3983	1555	756	690
Middle	Poor	Hispanic	6043	3751	772	697
Middle	Poor	Other	189	93	74	219
Middle	Poor	White	1209	556	452	485
Middle	Rich	Asian	721	1545	43	372
Middle	Rich	Black	432	1251	72	458
Middle	Rich	Hispanic	1035	1740	168	681
Middle	Rich	Other	122	415	34	157
Middle	Rich	White	3052	17340	2079	7203
Primary	MedPoor	Asian	1450	1204	277	232
Primary	MedPoor	Black	6378	5185	1708	2335
Primary	MedPoor	Hispanic	8027	7651	2754	2909
Primary	MedPoor	Other	851	839	602	814
Primary	MedPoor	White	7764	8505	9729	16436
Primary	MedRich	Asian	1250	1844	313	498
Primary	MedRich	Black	2609	3848	553	1694
Primary	MedRich	Hispanic	2980	5178	1200	2238
Primary	MedRich	Other	626	1108	363	647
Primary	MedRich	White	8924	17684	10377	21912
Primary	Poor	Asian	2167	677	79	117
Primary	Poor	Black	18604	5482	2174	2369
Primary	Poor	Hispanic	23350	12905	2740	2483
Primary	Poor	Other	956	469	330	753
Primary	Poor	White	5069	2438	1990	2907
Primary	Rich	Asian	2132	3619	113	937
Primary	Rich	Black	1184	2358	137	859
Primary	Rich	Hispanic	2793	3992	451	1769
Primary	Rich	Other	462	1259	83	454
Primary	Rich	White	8785	34068	3918	15929
Total Students in Sample: 484,020						

This sample was then used as a representative sample of the country and linked to each of the 36 state standards in order to determine how standards differed between grades and subjects. For instance, if the national sample of over 400,000 students were being assessed in Kentucky, Kentucky proficiency standards would be used to assess the proficiency rates at each grade. This process was repeated for every state, so that the observed proficiency rates of the random sample could be computed as though that sample were located and tested in each state.

In order to examine the impact of calibration, the random sample was then evaluated a second time for each state, using the most difficult proficiency standard for that state as the benchmark for all other grades. To illustrate, if grade seven reading had the hardest proficiency standard in Massachusetts (77th percentile, for example), then the proficiency rates for the other grades would be re-evaluated as if they, too, had a proficiency standard set at the 77th percentile. Doing so, we generated a figure for each state comparing the proficiency rates for each grade within the random sample under current proficiency standards to the proficiency rates that would be seen if the state used calibrated proficiency standards. The resulting data was used in the visualizations included in our Data Gallery.

KINGSBURY
CENTER AT NWEA