

**The State of State Standards:
Research Investigating Proficiency Levels in Fourteen States**

G. Gage Kingsbury, Allan Olson, John Cronin, Carl Hauser, Ron Houser

Northwest Evaluation Association

The passage of the No Child Left Behind Act (NCLB) by the federal government in 2002 provided public education in the United States with a number of vast opportunities as well as several daunting challenges. In reading the law and its associated regulations, it is not always clear where the challenges leave off and the opportunities begin. One of the primary effects of the act was to require each state to set proficiency levels in mathematics and reading and to categorize students into one of at least three proficiency categories, normally labeled *Basic*, *Proficient*, and *Advanced*. In this report we will most commonly refer to the level most states call *proficient* and will refer to minimum test score associated with this level as the *proficiency score*.

This legislation requires each state to set its own process for placing students into proficiency levels based on assessment results, but it does not specify how proficiency levels should be set. It is also the states' responsibility to define what a student's capabilities should be within each proficiency level. This provides states with a great deal of autonomy, but raises questions concerning the results of the process for setting proficiency levels.

With a couple of exceptions, all states have now accomplished this standard-setting task.. Now that the task has been accomplished, the following three questions are important given the latitude that the states have been given in setting proficiency levels:

1. **Are proficiency levels consistent and comparable among states?** While the federal law does not require this consistency, it is important to know if students and schools will be expected to perform similarly from one state to the next. Most prior studies have focused on this research question.
2. **Within each state, are proficiency levels consistent and comparable across grade levels in each subject?** Again, the law doesn't require this consistency, but it is required for teachers to know whether their students are well prepared for the next grade's challenges. Little prior research has addressed this question.
3. **Are proficiency levels consistent and comparable between subject areas within a state?** Since many funding decisions will be based on the percentage of students

attaining various proficiency levels, it is important to know whether the expectations for mathematics and reading achievement are similar. There is also little prior research addressing this question.

Some previous research concerning standard setting indicates that we should expect variability in standards as a result of the standard setters and the standard-setting process used (Jaeger, 1989; Reckase, 1998; Linn, 2003; Green, Trimble & Lewis, 2003). This may affect the consistency of standards within and among states, but the research does not identify the magnitude of the variability that might exist in an operational setting.

Two recent studies (McGlaughlin and Bandiera de Mello (2002); Linn, 2003) have investigated the variability in state proficiency standards by comparing student performance on state assessments to student performance on the National Assessment of Educational Progress (NAEP). These studies provide compelling evidence that performance measured by state proficiency levels differs markedly from performance measured by the NAEP exams. These studies suffer from three major limitations however. First, the NAEP examinations are not designed to directly align with the content standards of any one state. This non-alignment reduces the strength of the comparisons being made. Second, the NAEP examinations are no-stakes tests for the students and teachers involved. Individual student scores are not reported to parents or students and teachers do not receive classroom level results. The state assessment, however, typically has substantial stakes for both of these groups. This difference creates a risk that NAEP results may underestimate the actual level of student performance. Third, prior studies relied on extrapolations from group performance data to make comparisons of standards rather than collecting individual performance information from groups of students who had taken both their state test and NAEP.

To more fully address the three identified research questions, and to extend existing research, NWEA compiled a meta-analysis of 14 prior research studies investigating the student proficiency standards that have been established by various states¹. This paper synthesizes the results of those individual studies to address each of the three research questions described above.

¹ Arizona, California, Colorado, Iowa, Idaho, Illinois, Indiana, Minnesota, Montana, Oregon, South Carolina, Texas, Washington, Wyoming

This synthesis benefits from the characteristics of the NWEA measurement scales. The measurement scales, which we refer to as RIT scales, were designed to measure student performance and growth across grades (Ingebo, 1997). The scales were built to take advantage of strong measurement theory and experimental design, and have been demonstrated to be extremely stable, even over twenty years of development and use (Kingsbury, 2003). These design and stability of the scales allow results from different states to be examined on a single, common measurement scale. Thus, comparisons of cut scores may be made among states and within each state across grades or subject areas.

The synthesis also benefits from the unique characteristics of the NWEA assessment instruments which address some of the methodological limitations of prior studies. In each state, the NWEA instruments used were designed to align specifically with the content standards of the state. This designed content alignment enhances the quality of the score comparisons to be made in this paper. In addition to this content alignment, NWEA assessments are designed to match the capabilities of individual students. In both paper form (Achievement Level Tests) and computerized adaptive form (Measures of Academic Progress) NWEA assessments are designed to provide as much information about a student as possible, through the matching of test difficulty to student performance. Most school systems using NWEA assessments use their information to inform instructional decisions, inform placement, and assess growth. Attaching expectations to the assessment, communicating progress, and using results for instruction all help improve the level of student motivation and teacher interest in assuring students give their best effort. Finally, NWEA uses methods to identify scores of students who may be purely guessing or completing tests without serious effort. Districts typically retest these students to generate more valid score estimates.

While the issues concerning the comparison of proficiency levels are complex, this study advances our knowledge significantly. The purpose of this study is **not** to argue that some states have set their proficiency levels in error. Each state in this study has set standards with the intention that they be challenging and fair. All states used commonly accepted procedures to determine fair proficiency levels to associate with those tests. It is also not the intent of this paper to argue that proficiency levels need to be higher (or lower). No study, including this one, is capable of definitively identifying which state (if any) has determined the “right” proficiency levels. Rather, the study intends to extend our knowledge about differences and similarities in the

proficiency levels that states have set, and to identify the effects of inconsistency where it is identified.

Methodology

We studied fourteen states to generate this compilation. Each study statistically linked the minimum scores required for *basic*, *proficient*, and *advanced* levels of performance on each states' tests with the minimum scores associated with that performance level on a Northwest Evaluation Association exam that was aligned to the state's curriculum standards. All scores for the NWEA tests in a subject area reference a single, cross-grade, equal-interval scale developed using Item Response Theory methodology. These scales are referred to as RIT scales.

Data for the studies were collected between March, 1997 and May, 2003, and analyses were completed between May, 2003 and November, 2003. For purposes of this compilation, we validated that the scores the state used to define the performance levels during the original data collection were still in use on the date of the studies' release. Table 1 shows when data collection occurred for each state. At the time of this study, twelve of the states used assessments that were designed to align directly with their curriculum standards, while two states, Iowa and Montana, used the Iowa Test of Basic Skills, a nationally-normed assessment that reports student performance on a single set of curriculum standards that is not intentionally aligned to any particular state curriculum.

State and NWEA assessment records from over 136,000 students in 64 school districts were analyzed during the course of the 14 state studies. Alignment results for Idaho are not specified because NWEA provides the state assessment. As a result, no alignment between the NWEA test and state test is necessary. The state of Oregon licenses the NWEA scale and studies are periodically conducted to check alignment of the two scales. The most recent study was completed in 2001.

Table 1 – Data collection, district participation, and counts for the state alignment studies

State	Data Collected	Study Published	# of Districts Participating	# of students participating (reading)
Arizona	Spring 2002	July 2003	4	7,103
California	Spring 2003	July 2003	2	13,780
Colorado	Spring 2002	April 2003	11	46,198
Idaho	Northwest Evaluation Association provides the state assessment			
Illinois	Spring 2003	November 2003	8	5,235
Indiana	Fall 2002	August 2003	19	8,101
Iowa*	Fall 2001	August 2003	7	5,700
Minnesota	Spring 2003	In press	6	13,108
Montana*	Fall 2001	August 2003	7	5,700
Oregon	The Oregon Department of Education licenses the NWEA scale			
South Carolina	Spring 2002	April 2003	2	11,594
Texas	Spring 2003	September 2003	1	11,786
Washington*	Spring 1999	May 2000	3	4,814
Wyoming	Spring 2000	January 2001	4	3,110
Total			68	136,229

- Results for Iowa and Montana are based on an alignment study conducted on the Iowa Test of Basic Skills, a national test that is administered in the same form nationwide. Iowa and Montana currently use the ITBS for NCLB reporting.
- The Washington data were collected over a three year period between spring 1997 and spring 1999. The counts reported are for data collected in 1999.

In general, the same procedural and statistical methodology was used in each state to estimate their projected minimum proficiency score on the RIT scales. Appendix 1 documents instances in which the methodology employed substantively varied from our general procedures.

Most of the comparisons in this study were made using the level of performance defined as *proficient* for purposes of reporting related to the No Child Left Behind Act. In all states but Colorado, the level of performance defined as *proficient* for both state and NCLB reporting is identical. For Colorado, the level defined as *partially proficient* is used to define proficient performance for NCLB reporting. Because Colorado’s *partially proficient* standard is relatively low, readers should remember that Colorado uses a more stringent standard of proficiency for internal state reporting.

Our approach provides proficiency score estimates that can be directly compared across states within a subject area. Each state study generated these estimates from a sample that included a minimum of 1000 students who took both the mandated state test and a second NWEA test within

a month. In all states but Idaho and Montana, estimates were generated by direct assessment of students attending schools within the state.

Student records were matched using district assigned student ID numbers. All test records were screened to remove invalid scores. In all studies, more than 95% of the original sample of records were retained.

Content validity

The NWEA assessment used in each of the fourteen states was designed to directly align with each state's curriculum standards. This was accomplished by cross-referencing the state's content standards with the index that organizes the NWEA item bank, which contains more than 12,000 test items in reading and mathematics. All NWEA items are written in a multiple-choice format. The item format used by state assessments do vary, and while a few states use exclusively multiple-choice formats, most use a combination of multiple-choice, short answer, and extended response questions.

To statistically validate content alignment, concurrent and discriminant validity statistics were generated using simple bivariate Pearson correlation coefficients. In reading, the correlations between the NWEA assessment and the state reading or English/language arts tests ranged from .66 to .91. In mathematics, the correlations ranged from .69 to .92. Table 2 summarizes these correlations used to establish concurrent validity.

Of the 52 state, subject, and grade level correlations generated to document discriminant validity, only one state assessment (the South Carolina English/language arts assessment) at one grade correlated more closely with the NWEA math test than its reading counterpart. In general, same subject correlations were very high, with more than 80% of the correlations in reading producing a Pearson r above .75 and more than 75% of the correlations in mathematics above .80.

Table 2 - Summary of same subject Pearson correlation coefficients for state and Northwest Evaluation Association assessments at each grade level studied¹

	State reading/ELA to NWEA reading	State math to NWEA math
Below .70	3	1
.70 to .74	4	2
.75 to .79	19	8
.80 to .84	9	13
.85 to .89	6	13
Above .90	1	2

Linking state test scores to the RIT scale.

Three total score methods were used to estimate the minimum NWEA scale scores associated with each state assessment’s performance levels. The most straightforward was simple linear regression ($state_{pred}=a(RIT)+c$). Since we often observe departures from a linear relationship on the lower and upper ends of state test scales, a second order regression model was also employed ($state_{pred}=a(RIT^2)+b(RIT)+c$). For each of these methods, the RIT score was determined by substituting the appropriate state test proficiency score for $state_{pred}$ and solving the equation for RIT.

A fixed-parameter Rasch model (Ingebo, 1997) was the third method employed to estimate RIT performance level cut scores. In this method, each performance level on the state test was treated as a single test item. The assumption is that the performance level “item” should contain all the information about the difficulty of the test. Student abilities (RIT scores) were the “fixed parameter” used to anchor the difficulty estimate of the state-defined performance category to the RIT scale. The resulting difficulty estimate was taken as the RIT cut score for this method. This is referred to as the Rasch Status on Standard, or Rasch SOS, method. We have found that this method is often more accurate than conventional regression methods in estimating cut scores when they are near the low and high performance boundaries of a scale.

This is the first study to use three methods to estimate performance level scores. Employing three methods helps provide more reasonable estimates when the relationship between tests is not linear or when cut scores are near the high and low end of a scale.

¹ For most studies, the Pearson correlation was generated by subject and grade level. The Oregon study was not disaggregated by grade level.

Each method employed yielded an estimate of the state’s proficiency score for the subject area on the RIT scale. Selection of a single, best estimate for the subject area was made by examining and comparing prediction accuracy for each method.

Accuracy of Prediction

For 10 of the 14 state studies we were able to estimate the accuracy with which Northwest Evaluation Association assessments predicted proficiency status and performance levels on the state assessment. These estimates were not generated for the state’s of Montana and Iowa (because the study was conducted on the ITBS with Idaho data), for Oregon (where studies involved direct alignment of the scale) and Idaho (where NWEA provides the state test).

Table 3 shows the accuracy of prediction data for these state studies. We report the percentage of correct proficiency status predictions (proficient/not proficient) and the percentage of correct performance level predictions. We also report a prediction index statistic that shows the proportion of Type I errors relative to the number of correct predictions. For proficiency status, Type I errors occur when the NWEA assessment predicts a student will be proficient or better who in fact performs below proficient on the state test. For performance level, Type I errors occur when the NWEA assessment predicts a student will perform at a higher performance level than the student actually achieves on the state test.

$$\text{Prediction Index} = 1 - (\text{Type I Errors}/\text{Correct Predictions})$$

The higher the prediction index number, the larger the proportion of correct predictions for every Type I error. A prediction index statistic of .888 for example, indicates that there were about 9 correct predictions for every Type I error.

The use of the prediction index as an adjunct to a prediction accuracy statistic allows us to choose the method that is both highly accurate while simultaneously minimizing undesirable errors of overestimation (Type I errors). This better assures that the educators using the resultant cut scores receive projections that will not underestimate the number of students who may need added support to meet their state standards.

Table 3 – Accuracy of prediction estimates Northwest Evaluation Association Assessments with state assessments

State	Reading				Math			
	Proficiency Status		Performance Level Status		Proficiency Status		Performance Level Status	
	Pct. Correct	Pred. Index	Pct. Correct	Pred. Index	Pct. Correct	Pred. Index	Pct. Correct	Pred. Index
Arizona	84.0%	.911	63.0%	.756	85.7%	.919	65.2%	.726
California*	83.2%	.921			84.5%	.921		
Colorado	95.3%	.966	76.1%	.840	91.6%	.954	71.0%	.808
Illinois	88.6%	.928	69.5%	.804	93.1%	.945	71.1%	.888
Indiana	85.1%	.900	78.3%	.860	79.7%	.860	72.1%	.804
Minnesota**	90.4%	.944	59.0%	.633	89.4%	.935	57.2%	.611
South Carolina*	80.4%	.902	64.0%	.757	86.2%	.943	65.3%	.764
Texas	88.7%	.954	73.6%	.867	89.1%	.932	79.6%	.900
Washington	82.4%	.913			84.0%	.958		
Wyoming***	79.9%	.932			85.0%	.980		

* California and South Carolina combine reading and with language assessment

** Minnesota uses five performance levels. Data does not include grade 8 basic skills test

*** Performance Level Predictions were not generated for California, Wyoming, and Washington

In reading, the Northwest Evaluation Association assessments correctly predicted proficiency status on the corresponding state assessment in a range of 80% (Wyoming) to 95% of cases (Colorado). The accuracy of performance level predictions ranged from about 59% (Minnesota) to about 78% (Indiana). The prediction index statistics for reading that were based on proficiency status were all greater than .900. Prediction index statistics based performance level estimates were all greater than .750. In mathematics, prediction index statistics based on proficiency status were all above .910 with the exception of Indiana (.860). For performance level assignment, prediction index statistics were all greater than .600.

Given that so many factors may affect student performance on two tests (for example a student might be well rested and highly motivated on the day of the NWEA test and tired and hungry on the day of the state test), the results show that the NWEA assessments are quite effective in predicting student proficiency status and performance level. The accuracy of prediction did not generally seem to be compromised when students took a state assessment that combined reading and language.

The methods employed have resulted in estimated cut scores aligned on the RIT scale that predicted student performance with a high degree of accuracy. This allows for robust comparisons of the state standards studied using this scale. These comparisons were made using the performance level identified by the state as *proficient*. In all states except Colorado, *proficient* is the level used for Adequate Yearly Progress (AYP) reporting under the No Child Left Behind Act. Colorado uses the *partially proficient* level for this purpose. The percentile scores associated with all comparisons are based on Northwest Evaluation Association's most recent norm study (Northwest Evaluation Association, 2002). This study included more than 568,000 student reading test results and more than 578,000 mathematics test results for grades 2 through 10 collected during the spring of 2001.

Standards calibration

In order to answer the second research question it was necessary to estimate the calibration of standards. In some states a lack of consistency in the difficulty of the standard at different grade levels may pose problems. For purposes of this report, we define this problem as a lack of calibration.

A calibrated standard is one in which the minimum score required for proficiency at the exit grade is not substantively easier or more difficult than the standard at earlier grades.

We estimated calibration by finding the RIT and percentile score required to achieve proficiency on the 8th grade test and estimating, from that score, the percentile score and RIT that represented equivalent performance on the 3rd grade test. This was called the *calibrated score*. We subtracted the actual state proficiency score from the calibrated score to find the difference between them. A negative result would indicate that a state's 3rd grade state standard underestimates the score students need to achieve if they are going to be on track to meet the 8th grade standard. A positive difference indicates that the standard overestimates the score students will need to achieve in order to reach the standard.

We also used the calibrated score to estimate the percentage of students in our norm group whose 3rd grade performance would improperly classify them on the state test. This was done by taking the range of percentile scores associated with the RIT reflecting the actual 3rd grade standard and subtracting the percentile score associated with the 8th grade RIT reflecting the same standard. A

negative difference would show the percentage of students in NWEA's national norm population identified as proficient on the state's 3rd grade assessment who might not be proficient on the state's 8th grade test. Thus a negative result would indicate that the test errs on the side of identifying too many students as proficient (Type I error), while a positive result would indicate that the test errs on the side of identifying too few (Type II error).

Results

Are proficiency levels consistent and comparable among states?

The answer is no, although a pattern of greater consistency seems to be emerging. Prior studies have found large differences in state proficiency standards. We discovered similar differences in range, still we found that many state standards did cluster in a fairly narrow range of difficulty, particularly in reading, than we originally anticipated. In 3rd grade reading, for example, estimated minimum proficiency scores were calculated for 12 of the 14 states. These ranged from the 67th percentile (South Carolina, RIT 205) to the 13th percentile (Colorado & Texas, RIT 193). Seven of the twelve states, however, set proficiency cut scores in very a narrow range between the 31st and 33rd percentile (RIT 191 to 193). In 8th grade reading, estimated cut scores were also available for 12 states. These showed greater range with the highest cut score set at the 74th percentile (Wyoming, RIT 232) and the lowest at the 12th percentile (Colorado, RIT 206). Six states set their 8th grade standard near the low end of the range with proficiency scores ranging from the 32nd to the 38th percentile (RIT 218 to 221). The 35th percentile (RIT 219) represented the median proficiency standard.

Third grade math proficiency scores were estimated for 10 states. They ranged from the 75th percentile (South Carolina, RIT 208) to the 29th percentile (Illinois, RIT 193). Six states' scores clustered in a relatively narrow range between the 36th and 46th percentile (RIT 196 to RIT 199). The standard closest to the median was estimated at the 42nd percentile (RIT 198). Grade 8 proficiency scores were estimated for 12 states and they showed greater range, with the highest set at the 89th percentile (Wyoming, RIT 257) and the lowest at the 31st (Colorado, 225). Six state proficiency scores clustered between the 36th and 45th percentile (RIT 228 to 233). The standard closest to the median was estimated at the 42nd percentile (RIT 231).

Proficiency scores from seven states' high school exit examinations were also available. In reading, these ranged from the 77th percentile (Oregon, RIT 236) to the 14th percentile (California, RIT 208). For mathematics the scores ranged between the 73rd percentile (Washington, RIT 257) and the 13th percentile (California, RIT 232). Median cut scores were set at the 44rd percentile in reading (RIT 224) and at the 33rd percentile in mathematics (RIT 245). Table 4 summarizes the results of these comparisons.

Qualitative differences in the standards

Figure 1 (see page 14) helps illustrate some of the qualitative differences among state proficiency standards. An 8th grade student who can correctly answer items like sample question 1 would achieve the highest performance level in Arizona, Oregon, and Montana. That same student would not be assured of achieving the minimum proficiency score in Wyoming, a state in which students must perform at the 89th percentile equivalent to receive that designation. Conversely, a student who answers questions like sample question 2 correctly is assured of proficiency only in Montana. The level of performance reflected by that question would fall short of expectations in Wyoming, Arizona, and Oregon.

One can argue about the fairness of the Wyoming standard and the rigor of the Montana expectations, but it is clear that the differences among the standards reflect dramatically different visions of what it means to be proficient.

Table 4 - Cut scores representing “proficient” or “meets standards” level of performance on 14 state assessments

Reading

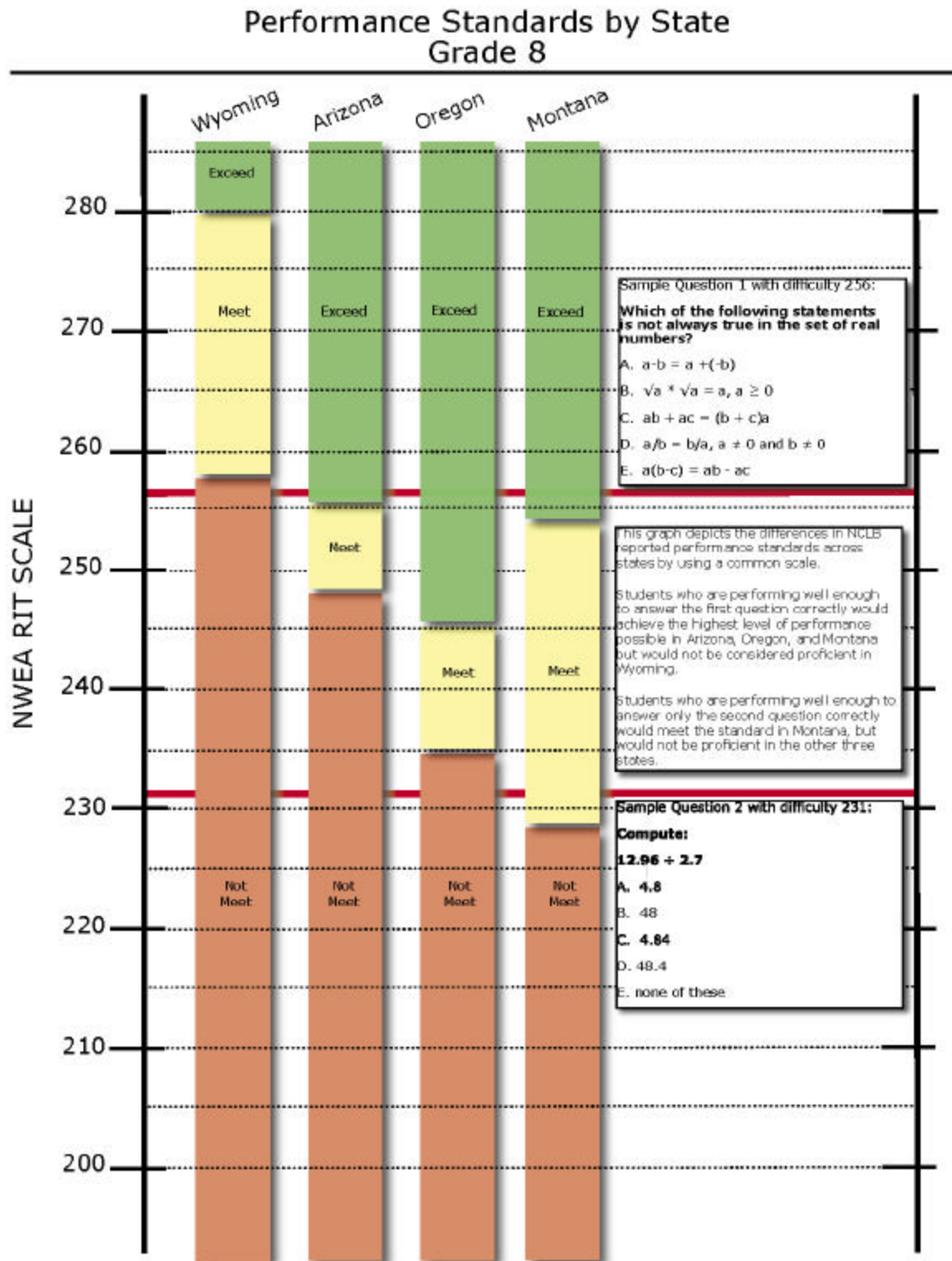
Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8			Grade 9			Grade 10		
State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile
SC	205	67	WY	214	73	SC	220	73	SC	221	63	SC	227	70	WY	232	74	MT	224	43	OR	236	77
CA	200	51	SC	213	70	CA	214	54	CA	216	46	WA	226	67	SC	230	68	IA	224	43	WA	227	51
MN	193	35	WA	207	53	AZ	210	45	MT	211	35	CA	221	50	OR	227	58	ID	221	37	ID	224	44
OR	193	35	CA	205	46	OR	209	42	ID	211	35	MT	218	43	CA	226	54	CO	204	9	MT	224	44
ID	193	35	ID	200	34	IL	207	37	IN	210	32	IA	216	37	AZ	224	49				IA	223	42
MT	193	35	MT	196	26	MT	206	35	IA	209	30	ID	215	35	IN	219	35				CO	209	15
IL	193	35	IA	196	26	ID	206	35	TX	208	28	TX	210	24	MT	219	35				CA	208	14
IN	192	32	CO	191	18	IA	205	32	CO	197	11	CO	206	18	IA	219	35						
IA	191	31				MN	204	30						ID	218	32							
AZ	190	29				TX	204	30						IL	218	32							
TX	179	13				CO	197	18						MN	218	32							
CO	179	13												CO	206	12							

Mathematics

Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8			Grade 9			Grade 10		
State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile
SC	208	75	WY	221	83	SC	227	76	SC	235	78	SC	242	78	WY	257	89	MT	242	47	WA	257	73
CA	204	60	WA	218	76	CA	225	70	CA	230	67	WA	242	78	SC	251	80	IA	241	44	MT	247	40
IN	201	50	SC	217	74	AZ	220	59	IN	221	47	CA	238	70	AZ	248	75	ID	240	42	IA	247	40
OR	199	46	CA	212	59	OR	215	46	ID	219	42	ID	225	44	CA	240	59	CO	235	32	OR	245	33
AZ	199	46	ID	205	39	ID	213	41	IA	218	40	MT	224	42	OR	235	50				ID	242	25
MN	198	42	IA	205	39	MT	212	38	MT	218	40	IA	222	38	ID	233	46				CO	233	14
MT	197	39	MT	205	39	IA	212	38	CO	207	19	TX	221	35	MN	231	42				CA	232	13
IA	197	39				MN	210	33				CO	216	26	IN	231	42						
ID	196	36				IL	210	33						IL	230	40							
IL	193	29				TX	209	31						MT	228	36							
						CO	201	15						IA	228	36							
														CO	225	31							

- Indiana tests students in the fall. Their cut scores were adjusted to reflect equivalent spring performance
- Colorado uses the partially proficient level of performance for NCLB reporting. To maintain consistency we report the level each state uses for NCLB reporting here.
 - The Texas estimate is based on the level for proficient performance that will be implemented in 2005.

Figure 1 – A comparison of performance levels for mathematics in four states



Are proficiency levels consistent and comparable within a state, within a subject area across grade levels?

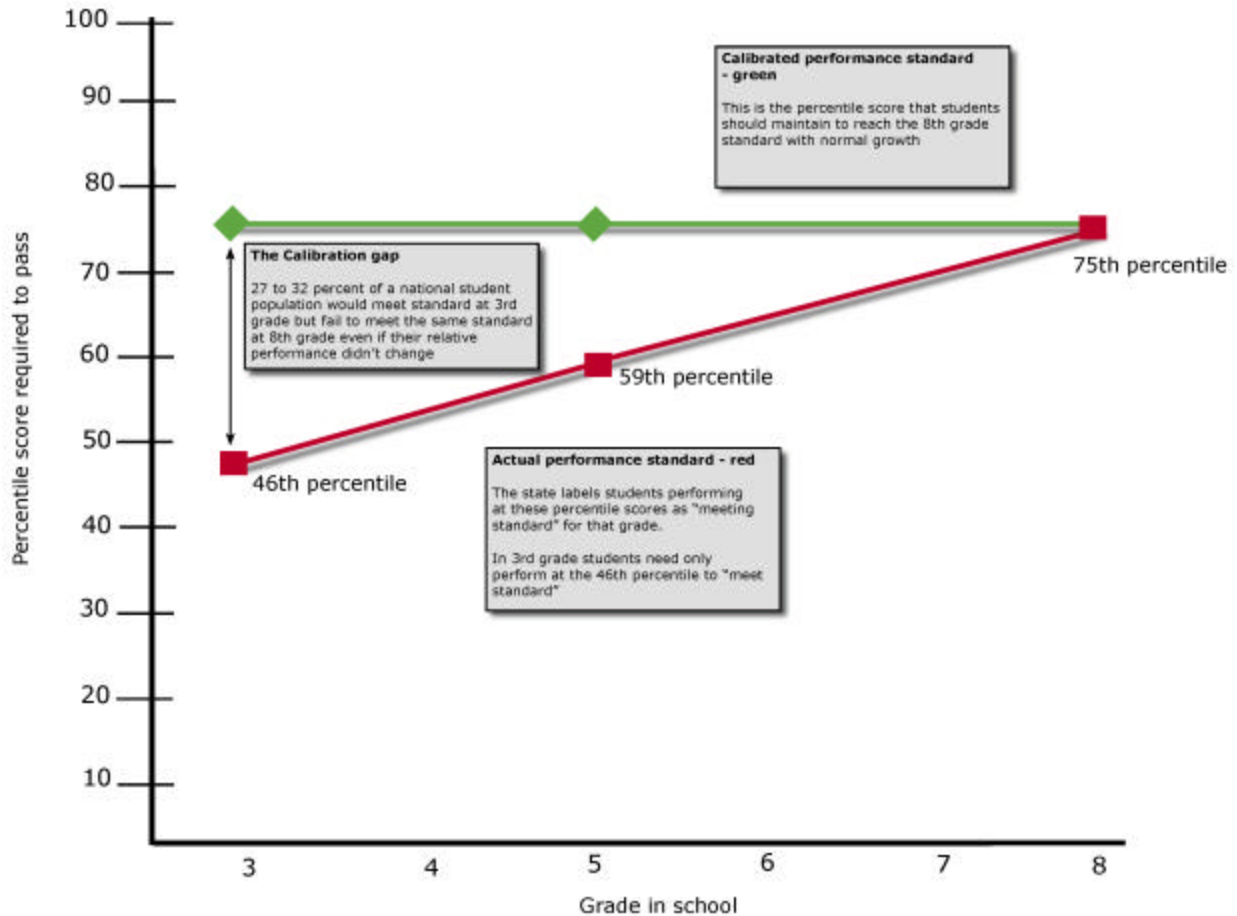
In some states lack of consistency between grade levels poses a serious problem that can be avoided if standards are calibrated. Standards that are not calibrated give students, parents, and educators an inaccurate perception about the child's standing relative to the expected level of performance. Students are reported as proficient in one grade who may not remain proficient in later grades even if they show normal growth.

Assume Xavier, for example, is a 3rd grader living with his family in Yuma, Arizona. Xavier scores at about the 46th percentile in mathematics on his state assessment, which is the minimum score for a rating of proficient, or *meets standard*. Xavier's teacher and parents believe he is performing at a level that is satisfactory relative to grade level standards.

Now let's move forward to 8th grade. Xavier takes the state level assessment again and achieves at the same level of performance, a 46th percentile score, relative to other students. This does not put Xavier anywhere close to the level required to meet the standard. Xavier's parents are alarmed that he is no longer meeting grade level standards and his teachers may come in for criticism because Xavier's performance "slipped". But Xavier's performance didn't slip. Instead he was the victim of a poorly calibrated standard that was too low at 3rd grade to reflect the performance that would be needed by 8th grade.

Figure 2 shows how this scenario, applied statewide, potentially affects many Arizona students. Indeed, the difference between the state's current 8th grade and 3rd grade standards may result in 27 to 32% of our national norm population being identified as proficient in third grade who might not achieve proficiency on the state's 8th grade assessment (Type I error). When calibrating a standard, it seems prudent to err, if one errs at all, on the side of committing Type II error, that is slightly overestimating the score students will need in grade 3 to meet the grade 8 standard. This helps assure that marginally qualified students are identified for the extra help they may need to stay above the standard.

Figure 2 – Example of a 3rd and 8th grade mathematics standard with poor calibration



In addition, an uncalibrated standard may leave the public with the impression that some grades within a school system are performing better than others. For example, if the third grade standard in reading is set at a much easier level than the 8th grade standard, the public may be led to believe that 3rd grade teachers are more effective than 8th grade teachers, when in fact no difference may really exist.

Tables 5 and 6 show the estimated calibration of the state tests studied to the 8th grade standard in reading and mathematics. Using Oregon’s reading assessment as an example, the state’s actual proficiency standard understates the performance needed to be on track for the 8th grade standard by 9 points on the RIT scale. As a result, somewhere between 22% and 24% of a national group of 3rd grade students who would be identified as proficient on Oregon’s test, would probably not achieve proficiency on the 8th grade test, even if they showed normal growth.

Most state reading assessments maintained a calibration of +3 to –3 points on our scale. The Oregon, Arizona, Texas and Washington tests each set standards for 7th or 8th grade demonstration of proficiency that were considerably higher than the required performance for 3rd or 4th grade.

Table 5 – Estimated calibration of state tests in reading

State	Actual grade 3-4 standard	Calibrated grade 3-4 standard	Actual grade 7-8 standard	Difference in scale score between actual and calibrated standard	Percent of students over/under estimated as proficient
Oregon	193	202	227	-9	22 to 24% overestimated
Arizona	190	199	224	-9	19 to 22% overestimated
Texas (gr 3 to 7)	179	188	210	-9	11 to 12% overestimated
Washington (gr 4 to 7)	207	212	226	-5	13 to 16% overestimated
South Carolina	205	208	232	-3	7 to 10% overestimated
Iowa	191	193	219	-2	3 to 5% overestimated
California	200	201	226	-1	1 to 5% overestimated
Indiana	192	193	219	-1	1 to 3% overestimated
Wyoming (gr 4 to 7)	214	215	232	-1	1 to 3% overestimated
Idaho	193	193	218	0	-1% to +1%
Montana	193	193	219	0	-1% to +1%
Colorado	179	178	206	+1	0 to 1% underestimated
Illinois	193	192	218	+1	1 to 3% underestimated
Minnesota	193	192	218	+1	1 to 3% underestimated

All but four of the 3rd grade state mathematics proficiency standards were calibrated to within +3 to –3 RIT points of 8th grade performance on our scale. We have already noted that Arizona’s assessment was out of alignment with the calibrated grade 3 standard falling 9 RIT points short of the actual score required to maintain proficiency at grade 8. Applied to NWEA’s norms, between 27 to 32% of the tested population may be identified as proficient in grade 3, but would not be performing at a level needed to achieve proficiency in grade 8. Our understanding is that the Arizona Department of Education recognizes this potential issue and intends to implement new mathematics proficiency scores for the 2004-5 school year. The Colorado, Illinois, and Wyoming grade 3 standards also overestimated the number of 3rd grade students who would be on track to meet the 8th grade standard.

Even modest gaps in calibration may sometimes lead to substantial differences in the number of students identified as proficient. In Idaho, for example, the actual 3rd grade standard is only 3 RIT points tougher than the calibrated 3rd grade standard. In spite of this, the standard in use overestimates the number of 3rd grade students who would be on track to meet the 8th grade standard by 7 to 11%.

Table 6 – Estimated calibration of state tests in mathematics

State	Actual grade 3-5 standard	Calibrated grade 3-5 standard	Actual grade 7-8 standard	Difference in scale score between actual and calibrated standard	Percent of students over/under estimated as proficient
Arizona	199	208	248	-9	27 to 32% overestimated
Colorado (gr5 to 8)	201	209	225	-8	15 to 16% overestimated
Illinois	193	198	230	-5	11 to 15% overestimated
Wyoming (gr4 to 7)	221	225	257	-4	6 to 7% overestimated
Idaho	196	199	233	-3	7 to 11% overestimated
Oregon	199	201	235	2	4 to 9% overestimated
South Carolina	208	210	251	-2	3 to 7% overestimated
Texas (gr 5 to 7)	209	211	221	-2	3 to 6% overestimated
Washington (gr 4 to 7)	218	219	242	-1	1 to 3% overestimated
Minnesota	198	198	231	0	-1 to +1
Iowa	197	196	228	+1	1 to 4% underestimated
Montana	197	196	228	+1	1 to 4% underestimated
California	204	203	240	+1	1 to 6% underestimated
Indiana	201	198	231	+3	8 to 12% underestimated

Are proficiency levels consistent and comparable within a state, between subject areas?

No. In all states but Oregon, the proficiency standard set for mathematics was more difficult than the standard for reading. In a few states the difference was very large. In Arizona, the most obvious example, the mathematics proficiency score would not be achieved by about 75% of our norm group, while only 47% would not achieve the minimum score in reading (see Table 7).

Table 7 – Differences in reading and mathematics percentile cut scores for the proficient level of performance – 7th or 8th grade

State	Math cut score (percentile)	Reading cut score (percentile)	Difference
Arizona	75	47	28
Colorado	31	12	19
Wyoming	89	74	15
South Carolina	80	68	12
Idaho	46	32	14
Washington	78	67	11
Texas	35	24	11
Minnesota	42	32	10
Illinois	40	32	8
Indiana	42	35	7
California	59	54	5
Montana	36	35	1
Iowa	36	35	1
Oregon	50	58	-8

These differences do reflect concrete distinctions in what is actually expected of students. For example, this question reflects Idaho’s 8th grade reading proficiency standards (RIT 218, 32nd percentile) and students performing at the minimum score would answer it correctly 50% of the time:

Which sentence implies a cause-and-effect relationship?

1. When you spray fruit trees, use a nontoxic spray.
2. If you wish to have worm-free fruit, spray your trees.
3. Don’t damage the roots which grow near the surface.
4. Plant fruit trees in winter when they are dormant.

A clear majority of 8th grade students nationally easily answer this question. Now consider an example of a question with a difficulty that approximates the minimum proficiency score for Idaho’s 8th grade math standards (RIT 233, 46th percentile).

The volume of a rectangular solid is 480 cubic inches.

If the length is 10 inches and width is 8 inches, how high is the solid? (A box is the shape of a rectangular solid)

1. 6 inches
2. 8 inches
3. 10 inches
4. 12 inches
5. 20 inches

Only about half of 8th grade students nationally answer this question correctly and the item aptly shows that the differences between Idaho's reading and math standards reflect true differences in the difficulty of content.

Wyoming's 8th grade reading and mathematics proficiency standards are the most rigorous of those we reviewed for this study and provide additional contrast. The reading standards require roughly the vocabulary and reading skill needed to answer the question¹ below (RIT 232, 74th percentile) correctly 50% of the time:

Having no representatives in the state legislature, black people had to rely on white legislators willing to introduce bills on their behalf, on lobbying efforts, and by uniting in political groups and coalitions that voiced the collective needs of the black community. This took place in a time when many black men were employed on the railroad and often out of town, or working in menial jobs, struggling to provide their families with the necessities of life; much of the work was carried out by a handful of people, as paid volunteers.

This paragraph emphasizes

1. the objectives of the African-American community during the early 1900s.
2. the reasons why it was hard for African-Americans to influence legislation.
3. the types of employment available to African-Americans.
4. the process for how a bill becomes law

This item is challenging for most 8th graders (72% of these students will miss this item 50% of the time), but the average 8th grader is probably not intimidated by it. Contrast this with a mathematics question that approximates the expected proficiency demonstration (RIT 257, 89th percentile):

¹ Copyright 1980 by the Georgian Press Company, Portland, Oregon. Used with permission of the publisher.

One airplane is located 200 km north and 50 km east of an airport. A second plane at the same altitude is located 30 km north and 100 km west. The distance between the planes is closest to

1. 150 km
2. 200 km
3. 300 km
4. 350 km
5. 400 km

This item would intimidate many adults and 89% of 8th graders miss this item 50% of the time. While the difficulty of Wyoming's standard is atypical, the dramatic contrast in difficulty between a challenging reading and challenging mathematics standard is easily apparent.

Discussion

Implications of applying the NCLB accountability rules to widely differing state standards.

The No Child Left Behind Act was intended to ensure that each and every student leaves school with the basic skills needed to be successful as workers and citizens. The Act also contains escalating sanctions that are imposed on schools failing to make Adequate Yearly Progress targets.

The Act leaves it to states to determine the level of performance that reflects adequate preparation or proficiency. From the evidence we and others have collected so far, it is clear that there are very large differences in the difficulty of the various state proficiency standards. In states such as Wyoming the standards are set at levels that may be beyond what is needed for a student to have good prospects for completing college. The U.S. Census Bureau (2002) reports that about 38% of the adult population between the ages of 25 and 29 have completed a minimum of an occupational or academic associates degree or better and that about 29% of the same population have completed a bachelors degree or better. Wyoming's standards, which are set at NWEA's 74th percentile in reading and the 89th in math are high enough that there is little question that some students identified as not proficient on their 8th grade state tests will go on to complete college degrees. Wyoming is not alone in this. The mathematics standards currently in place for 8th grade in Arizona and South Carolina are similarly high.

Minimum proficiency standards that are set higher than the level needed to pursue baccalaureate education may create misconceptions about the effectiveness of schools among the public. When newspapers report that 75% of their states' 8th graders failed to meet a proficiency standard, readers would naturally assume that middle schools have failed to fulfill their mission. We wonder how many reports may have been written about the "crisis in middle school mathematics education" in states like

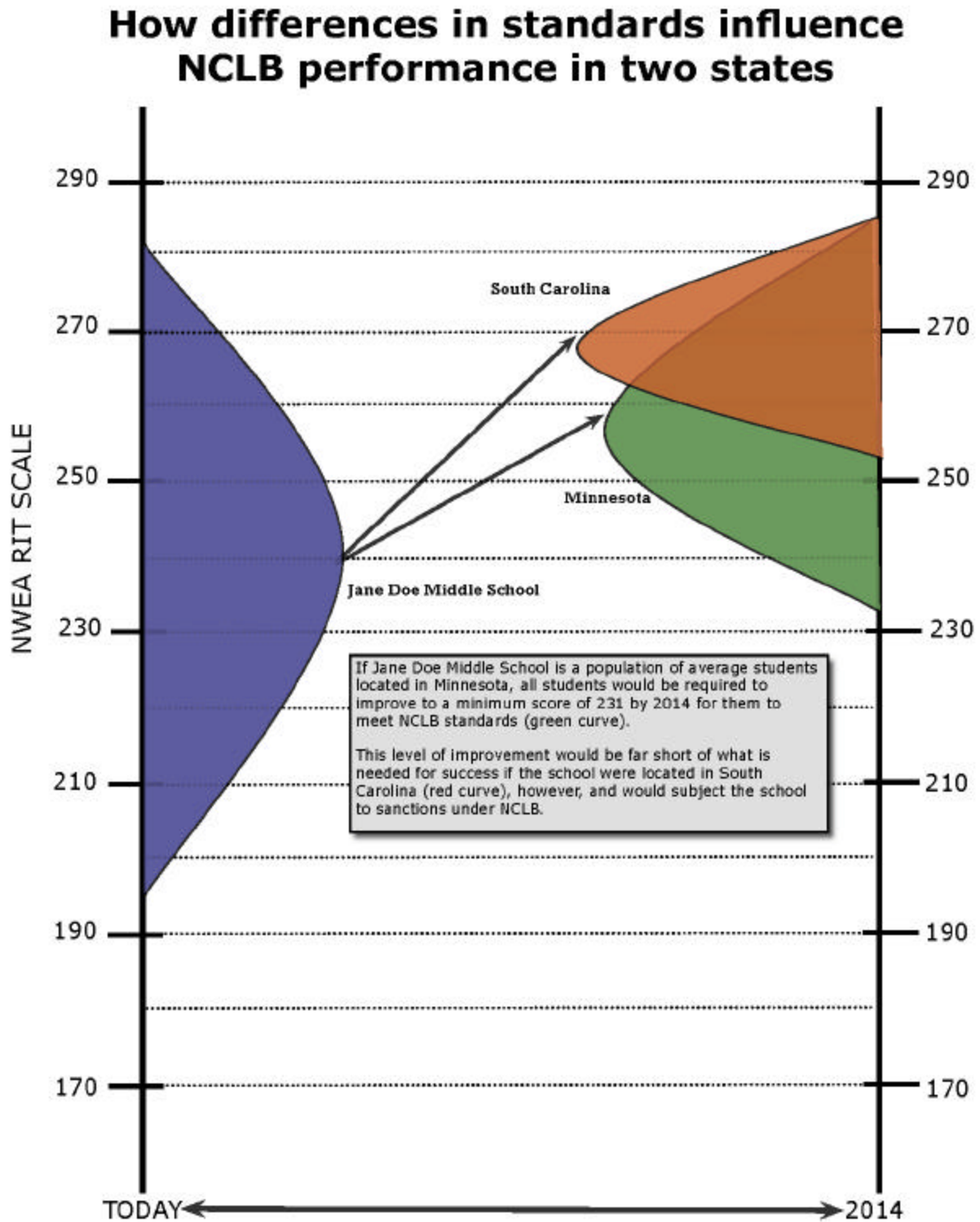
Wyoming, Arizona, Washington, and South Carolina, when the standards employed reflect expectations that are beyond those needed by college bound students. And national perceptions about the presumed failures of our nation's schools have been strongly influenced by NAEP proficiency standards that are set at a level not regularly attained by most students in schools anywhere in the world (Linn, 2003). One way to correct this problem is to find better ways to measure and communicate both student performance on the standard and the standard's difficulty to the public.

The range in difficulty among state standards also affects the prospects for schools to make the required Adequate Yearly Progress toward NCLB goals. Linn (2003) argues persuasively that virtually none of the past performance data available from state tests shows that a significant number of schools have ever achieved the level of improvement needed to assure all students achieve most state proficiency cut scores by 2014. It also is quite clear that those states with proficiency standards above the 70th percentile will not have the same prospects for achieving this goal as those that have set their targets below the 33rd percentile. This creates a problem when the accountability sanctions are uniform, as they are under NCLB.

Consider this example. Assume that Jane Doe Middle School is a school in Minnesota whose performance and distribution reflects our 8th grade averages (see figure 3). To meet the requirements of NCLB, which mandates that all students achieve proficient performance by 2014, Wilson has to improve the performance of each and every student to 231, a score representing the 42nd percentile. Assume further that Wilson accomplishes this very tall order. Not only will this school avoid NCLB sanctions, it will be celebrated (and should be) as a school that has remarkably improved the future prospects for its children.

Now let us add a bit of drama. Assume that a tornado tossed Doe Middle School and all of its educators, students and families onto a parcel of land in South Carolina in 2002. Despite the trauma of relocation, Doe students achieve the same improvement in twelve years. Unfortunately, because of the tornado's unfortunate choice of drop-off location, our school fails, miserably, to meet NCLB requirements. NCLB sanctions are imposed that include required opportunities for open enrollment of students with the district required to shoulder the financial burden of providing transportation; requirement for direct support for tutoring and other supplemental services; and possible dismissal and reconstitution of school staff (U.S. Department of Education, 2003). The only reason for the different treatment is that the tornado deposited our school in a state with standards reflecting

Figure 3 – Illustration of the Improvement Required to Meet NCLB Requirements in Two States



higher aspirations. One unfortunate consequence of NCLB is that it penalizes states with high standards while rewarding states that have less high standards. That said, Minnesota's standards seem quite reasonable and reflect ambitious expectations when applied to all students.

The need to improve calibration of proficiency standards within states.

If we are truly to leave no child behind, state standards and assessments must accurately identify all of the students in need of support in the early grades. There is an unfortunate tendency for 3rd grade standards to understate the level of performance needed to pass proficiency standards in the upper grades. This leads to three problems for schools and communities.

First, it creates a very serious problem for teachers and families who assume that children are performing and improving as expected when they are not performing at a level that will make them successful against the standards in the future. These students may not receive the additional help and services that would be needed to meet the more difficult proficiency standards that are applied in future grades. In a few states, this can impede their prospects for promotion to other grades or graduation.

Second, it creates potential for misallocation of public resources. When standards are poorly calibrated, regulators and lawmakers can too easily assume that elementary schools perform well relative to the rest of the school system. This assumption might lead them to make decisions that reduce funding and support to elementary programs in an effort to bolster support for the upper grades. The ultimate result may actually be the opposite of what was intended. Elementary students who were not adequately prepared for the more difficult middle level standards may not receive resources and support to improve their chances, making them even less prepared when they arrive in middle school classrooms.

Third, poor calibration may create public misconceptions about educators' performance. When twice as many 3rd graders pass a standard as 8th graders, the press and public will reasonably assume something is amiss with the 8th grade program. After all most reasonable persons would assume that the 8th grade standard and 3rd grade standard were set at the same relative difficulty. When the perceived gap is a product of differences in the standard and not performance, however, these assumptions can create unexpected problems. Articles in the press appear criticizing the performance of 8th grade teachers. Expert 8th grade teachers seek to reduce stress by transferring to grade levels that do not require high stakes testing (grade 9 for example). Desirable job candidates turn down 4th grade job offers in favor of the less stressful opportunities offered by teaching first grade, another grade that is not required to test.

In the meantime, articles in the press appear speculating about the reasons why elementary schools are so successful when middle schools are such horrible failures. While it would be speculation on our part to suggest that the more difficult standards often being applied at secondary schools may have already contributed to what seems to be an ensconced cultural perception of secondary school failure, measurement experts do have an obligation to the public to ensure that poor practice does not create or compound misconceptions about schools.

The need for dialogue about the differences in difficulty between math and reading standards.

The math standards in almost every state are more difficult than their reading standards. For the most part educators seem to assume that this is as it should be, perhaps because of criticism in the popular press about the quality of math and science preparation in our schools. Research as far back as Jaeger, Cole, Irwin, and Pratt (1980), however, found that standard-setting procedures may result in setting math standards that are more difficult and variable than corresponding reading standard. Green, Trimble, and Lewis (2003) reported a more recent study of multiple standard-setting procedures that were employed by the Kentucky Department of Education (2001). This work provided evidence that conventional book-marking procedures for setting standards, the method most commonly used by states, produced math standards that were substantively more difficult than reading standards.

It may be that more difficult math standards are justified. But, given the vast differences in results produced by different standard-setting procedures, it seems sensible to try to better understand the underlying psychology of these processes and reasons for well-documented inconsistencies before embracing that conclusion. In most states, groups of teachers have played a critical role in determining standards. Usually, teachers who are subject matter experts are chosen to participate in these processes. Researchers should also study the dynamics of these processes in order to explore what role differences in culture between math and reading teachers might play in the results. For example, many teachers with expertise in K-12 reading employ their expertise by working with the struggling and emerging readers in their schools. Math teachers who are perceived as expert, on the other hand, more often teach elite students in the upper grades. These and other fairly obvious factors (including perhaps differences in the college grades of math and reading teachers) may lead to profoundly different dynamics and results during the standard-setting process. Of course, it is not clear that this is actually the case, but one wonders what happened when some states set math standards that cannot currently be attained by most college bound students and a reading standard that is below the 50th percentile.

There are consequences attached to these differences that may lead to misallocation of resources. When a state's math standard is far more difficult than standards set in other subjects, huge changes in the allocation of teacher time, supplemental resources, and cash may be needed if a state is serious about all students attaining them. These resources will not generally come from increased investment in education by the community; at least some of the resources for mathematics will be cannibalized from other sources. Do we compromise programs like reading support for English language learners, high quality science or social studies instruction, or our music and physical education programs because math is THAT important? When the minimum standard for mathematics is 15 to 40 percentile points more difficult than reading, as it is in several states, this seems like a reasonable question to ask.

Conclusion

While the detailed results vary from one state to the next, examination of the results from all 14 states leads us to three general conclusions:

- Proficiency standards among states differ enough to cause substantial differences in the percentage of students categorized as proficient from one state to another, even if the students have exactly the same skills.
- Proficiency standards within individual states differ across grades enough that they may provide teachers with inconsistent proficiency indications for a large percentage of students.
- Proficiency standards between subject areas within and across states differ enough that they may provide schools with inconsistent information when comparing proficiency of students in reading to proficiency of students in mathematics.

Students and educators deserve to know what is expected of them, and states' efforts to set content standards and standards of performance have clearly helped schools bring greater focus to improving achievement. Pursuit of improvement requires that public policy, resources, and sanctions be applied in a rational, purposeful, and prudent fashion. The differences among state standards and the lack of calibration of standards across grades revealed by this study may result in policies, sanctions, and resource allocation that are counterproductive to the ultimate goal of high achievement for all students.

It is not rational to reconstitute some schools because all students didn't get to the 80th percentile in mathematics and simultaneously reward schools in other states that are barely making Adequate Yearly Progress toward a standard set at the 20th percentile. We do not act purposefully if we set artificially low standards at third grade that prevent us from identifying students who need help to reach the standard at eighth. It does not seem prudent to expect students to achieve math performance that college students might find challenging while setting standards in reading that are not equivalent to the level of performance required to read a daily newspaper.

Education trends in the United States sometimes swing like a pendulum, with programs and instruction that are in favor this year, and out of style five years later. NCLB provides important information that may be used in stopping the pendulum, by offering evidence of two kinds. First, it requires that schools have evidence concerning the percentage of students identified as having "proficient" status. Second, it requires that this percentage change year by year, so that every student in every subgroup is eventually identified as "proficient". The NCLB metrics reward and punish schools by determining whether they make adequate progress toward turning the non-proficient into proficient. Unfortunately this metric is not entirely adequate. There is no reward turning proficient students into accomplished ones. Many students perform too far below the proficiency bar to change their status in the short term. The NCLB metric provides incentive for schools to focus on kids who can change their proficiency status this year at the expense of these kinds of students.

The key element that is not represented in NCLB metrics is **individual growth**. A better accountability system would reward schools for the growth they nurture in students. Producing more growth in high achieving students would be recognized. Moving low achieving students closer to success would be noted. Growth is an element that is more difficult to measure, but crucial to achieving genuine equity. The proficiency standards discussed in this study are useful in measuring status and change, but they create a subtle inequity by focusing schools on the relatively small number of students who are nearly proficient, and diverting their attention from those who are far from proficient, or accomplished. The current NCLB legislation is a positive step, but won't be useful to teachers until the growth of each and every student is measured as an indicator of success.

We have set the right course by insisting that schools work to ensure all students achieve standards that enhance their opportunities for success in life. As we embark on that course, let us be flexible enough to adapt approaches to their implementation and enforcement that assure we reach our intended destination.

References

Green, D.R., Trimble, C. & Lewis, D.M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice*, No. 1, 22-32.

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

Jaeger, R. M. (1989). Certification of student competence. In R.L. Linn (ed.), *Educational Measurement* (3rd ed., pp.485-514). New York: Macmillan.

Jaeger, R. M., Cole J., Irwin, D.M., and Pratt, D.J. (1980). *An interactive structure judgment process for setting passing scores on competency tests applied to the North Carolina high school competency tests in reading and mathematics*. Greensboro, NC: Center for Education Research and Evaluation, University of North Carolina at Greensboro.

Kentucky Department of Education. (2001). *Standard setting: Synthesis of three procedures and findings*. Frankfurt, KY: Kentucky Department of Education. Available at www.kde.state.ky.us.

Kingsbury, G. (2003, April). *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.

Linn, R. (2003, July). *Accountability: Responsibility and Reasonable Expectations*. CSE Report 601. CRESST/University of Colorado at Boulder.

McGlaughlin, D. & Bandiera de Mello, V (2002, April). *Comparison of state elementary school mathematics achievement standards using NAEP 2000*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Northwest Evaluation Association (2002). *RIT Scale Norm*. Portland, OR.

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practices*, 17(2), 13-16.

U.S. Census Bureau (2002, March). Educational Attainment of the Population 15 Years and Over, by Age, Sex, Race, and Hispanic Origin: <http://www.census.gov/population/www/socdemo/education/pp1-169.html>. Page accessed on November 13, 2003.

U.S. Department of Education (2003, November). *Stronger Accountability: Questions and Answers on No Child Left Behind*. <http://www.ed.gov/nclb/accountability/schools/accountability.html#5>. Page accessed on November 13, 2003

Appendix 1 - Notes on the Included State Studies

Arizona

Data were collected for the standards in effect during the 2001-2002 school year. Arizona will implement recalibrated scales in 2005, in an effort to improve the consistency among their minimum performance level scores from grade to grade. In Arizona, the 8th grade math cut score we recommended is not the score that produced the highest prediction index number because of an excessively high number of Type II errors generated by that method. The alternate method predicted with about 85% accuracy with many fewer Type I errors.

Arizona reports four performance levels instead of the NCLB minimum of three.

California

Data were collected for the standards in effect during the 2002-2003 school year. Our estimates for California are based on two studies. Each involved samples of students taken in one district. A large school district was used to make the proficiency score estimates for grades 3 through 8 and a second large, high school district in a different part of the state was used for the 10th grade estimates.

The accuracy of performance level assignment was not evaluated in our California studies. We employed two linear and second-order regression methodologies to estimate cut scores and did not employ Rasch methodologies in these two studies.

California combines their reading and language arts assessments into one reported score. The NWEA reading assessment is not designed to have close content alignment with some of the language arts portions of this assessment. In the three states that combined reading with language arts, NWEA tests had slightly lower, but still satisfactory, rates of accurate prediction for proficiency status.

Colorado

Data were collected for the standards in effect during the 2001-2002 school year. At the time of the study Colorado did not test students in mathematics prior to grade 5. Colorado also reports four performance levels instead of the NCLB required three.

Colorado is the only state we studied that uses the minimum score they have defined as *partially proficient* to represent *proficient* performance for NCLB purposes. We used their NCLB level to maintain consistency in our meta-analysis of state standards. Had we used the level the state has formally defined as *proficient*, the Colorado standards would range from the 35th to the 41st percentile in reading and the 47th to the 74th percentile in mathematics.

Idaho

Northwest Evaluation Association provides the state test so no state alignment study was conducted. The standards in effect were those for the 2002-2003 school year.

Illinois

Data were collected for the standards in effect during the 2002-2003 school year. Illinois reports four performance levels instead of the NCLB required three.

Indiana

Data were collected for the standards in effect during the 2002-2003 school year. Indiana combines their reading and language arts assessments into one reported score. The NWEA reading assessment is not designed to have close content alignment some of the language arts portions of this assessment. In the three states that combined reading with language arts, NWEA tests had slightly lower, but still satisfactory, rates of accurate prediction for proficiency status.

Indiana administers their state test in the fall, so the estimated cut scores in Table 4 have been adjusted to reflect spring performance. To make the adjustment we added the rounded fall to spring growth norm (NWEA tests are normed for both performance and growth) for the grade that appeared our 2002 study to the fall score.

Iowa

The standards that were in effect for the 2002-2003 school year were used. Iowa has no statewide curriculum standards and no official state test. For purposes of NCLB required reporting, all districts in the state use the Iowa Test of Basic Skills in grades 4, 8, and 11. The scores associated with the proficiency levels reported by ITBS are those used to determine our estimates. Because equivalent versions of the ITBS are used nationwide, alignment of cut scores was estimated using a population from Idaho who took a version of the ITBS that is equivalent to the one used in Iowa.

Minnesota

Data were collected for the standards in effect during the 2002-2003 school year. Minnesota administers the Minnesota Comprehensive Assessment (MCA) at grades 3 and 5. This assessment reports five levels of performance (all other states reported three or four). This is almost certainly one reason why the accuracy of prediction for performance levels on the Minnesota test was lower than the accuracy levels seen in the other state studies. At grade 8, Minnesota administers the Basic Skills Test (BST). Results in this test are reported as pass/fail only.

Montana

The standards that were in effect for the 2002-2003 school year were used. Montana will implement a new statewide assessment during the 2004-2005 school year. At the time of the study Montana used a nationally-normed assessment, the Iowa Test of Basic Skills, for NCLB reporting. Because equivalent versions of the ITBS are used nationwide, alignment of cut scores was estimated using a population from Idaho who took a version of the ITBS that is equivalent to the one used in Montana. The state intends to fully implement an assessment aligned with curriculum standards during the 2004-2005 school year.

Oregon

The standards that were in effect for the 2002-2003 school year were used. Oregon licenses the NWEA scale so scale alignment estimates are determined by periodic checks of item calibrations between the two scales. The most recent calibration was conducted during 2001.

South Carolina

Data were collected for the standards that were in effect during the 2002-2003 school year. South Carolina uses a combined reading/language arts assessment. The NWEA reading assessment is not designed to have close content alignment with some of the language arts portions of this assessment. In the three states that combined reading with language arts, NWEA tests had slightly lower, but still satisfactory, rates of accurate prediction for proficiency status. South Carolina reports four performance levels instead of the NCLB required three.

Texas

The Texas Assessment of Knowledge and Skills (TAKS) was implemented as the state assessment in 2002-2003. The cut score in use for 2002-2003 was an interim cut score that is two standard errors of measure below the score that will be phased-in for the 2004-2005 school year. This cut score is phased in and a minimum proficiency scale score that will go into full effect in for the 2004-2005 school year. The cut score will be one standard error of measure below the eventual standard for 2003-2004. For purposes of this study we used the standard that will be in effect during 2004-2005. This is the highest or most challenging cut score.

Data for this study were collected from the 2002-2003 administration of TAKS conducted by a single large school district in the state. Because this district does not administer NWEA tests at all grades in which TAKS is used, we only reported estimated cut scores for those grades in which the district used both tests.

Washington

Data were collected for the standards that were in effect during the 1998-1999 school year. State required tests are at grades 4, 7, and 10. Washington reports four performance levels rather than the NCLB required three.

Wyoming

Data were collected for the standards that were in effect during the 1999-2000 school year. At the time, state required tests were mandated at grades 4 and 8. Performance levels were not evaluated during this study.