A Study of the Ongoing Alignment of the NWEA RIT Scale with the Arizona Instrument to Measure Standards (AIMS)

John Cronin and Branin Bowe

October 26, 2005

Northwest Evaluation Association
5885 SW Meadows Road, Suite 200
Lake Oswego, OR 97035-3526

www.nwea.org
Tel 503-624-1951
Fax 503-639-7873

A Study of the Ongoing Alignment of the NWEA RIT Scale with the Arizona Instrument to Measure Standards (AIMS)

John Cronin and Branin Bowe

September, 2005

Each spring, Arizona students participate in testing as part of the state's assessment program. Elementary and middle school students in grades 3 through 8 take the Arizona Instrument to Measure Standards – Dual Purpose Assessment (AIMS DPA) in reading, writing, and mathematics. These tests serve as an important measure of student achievement for the state's accountability system. Results from these assessments are used to make state-level decisions concerning education, to meet *Adequate Yearly Progress* (AYP) reporting requirements of the *No Child Left Behind Act* (NCLB), calculate status and improvement indicators for AZ LEARNS, the state accountability system, and to inform schools and school districts of their performance. The Arizona Department of Education has developed scales that are used to assign students to one of four performance levels on these tests.

Many students who attend school in Arizona also take tests developed in cooperation with the Northwest Evaluation Association (NWEA). The content of these tests are aligned with the Arizona standards and they report student performance on a single, cross-grade scale, which NWEA calls the RIT scale. This scale was developed using Rasch scaling methodologies. RIT-based tests are used to inform a variety of educational decisions at the district, school, and classroom level. They are also used to monitor the academic growth of students and cohorts. Districts choose whether to include these assessments in their local assessment programs. They are not state mandated.

In order to use the two testing systems to support each other, an alignment of the scores from the state and RIT-based tests is as important as curriculum alignment. A July, 2003 study first established estimated RIT scores that aligned with the equivalent cut points on the AIMS scale (Cronin, 2003). Because Arizona expanded the number of grades tested in spring of 2005, we undertook a study to estimate the aligned cut scores for the grades added and attempted to determine whether previous estimates of cut scores had changed. We estimated the relative accuracy with which the NWEA assessments continued to predict AIMS results. Finally, we developed estimates for both the spring and prior fall RIT scores so schools may use spring results to assess their students' likelihood of success on AIMS. The primary questions addressed in this study are:

- What RIT scores correspond to various performance levels on the AIMS tests?
- How do these RIT scores differ from the 2003 estimates of performance levels?
- How well can performance on the Arizona assessments be predicted from RIT scores when NWEA assessments are administered in the same fall and the prior spring?

Method

Our study included test records from over 15,000 students enrolled in 3 Arizona school systems. These students had taken both the state assessment and NWEA assessments in spring of 2005; many had also taken NWEA assessments in fall of 2004. Student records were included when a student had both a valid NWEA scale score and a valid AIMS score in the equivalent subject for the spring season. We excluded records in which students had been given accommodations on the state assessment.

The methodology used to complete this validation study was identical to that used in almost all of the state studies that we have completed in recent years (see Kingsbury et al, 2003). To conserve space, we refer readers to this study, "The State of State Standards", which is available on our website (www.nwea.org/research/national.asp) , for more detail about the methods we use to conduct scale alignment studies.

Results

Descriptive Statistics

Table 1 reviews descriptive statistics for the AIMS and NWEA assessments. The median RIT scores for this sample in reading and mathematics were generally 0 to 2 points below the median for the 2005 NWEA norm population sample. The distributions in both subjects showed some evidence of a negative skew. Nevertheless, the sample provides reasonable numbers of students who perform at all levels on the test scales and this assures that the statistical methods applied have an adequately large sample to derive good estimates of performance levels that are at the higher and lower ends of a test scale.

Pearson correlations

Table 2 shows the Pearson correlations for each grade. Concurrent validity was tested by examining same subject Pearson correlations between the NWEA and AIMS assessments. AIMS reading to NWEA reading coefficients for tests administered during the same season were very high, ranging between .79 and .85. For NWEA tests administered the prior fall, correlation coefficients ranged between .78 and .83, Correlations between the two NWEA reading assessment terms were slightly stronger, as expected, with coefficients ranging between .84 and .87. In mathematics, correlations between the AIMS and the NWEA mathematics assessments were also very high. Correlations between AIMS and NWEA mathematics ranged from .84 to .88 for both same season administrations and between .81 and .86 for NWEA administrations that occurred during the prior fall. Correlations between the spring and prior fall NWEA mathematics assessments ranged between .83 to .89. The strength of the correlations among the assessments suggests that tests are measuring the same general constructs and that the relationship maintains reliable when the tests are separated by time.

Table 1 – Means, Standard Deviations, and Medians for AIMS and NWEA assessments

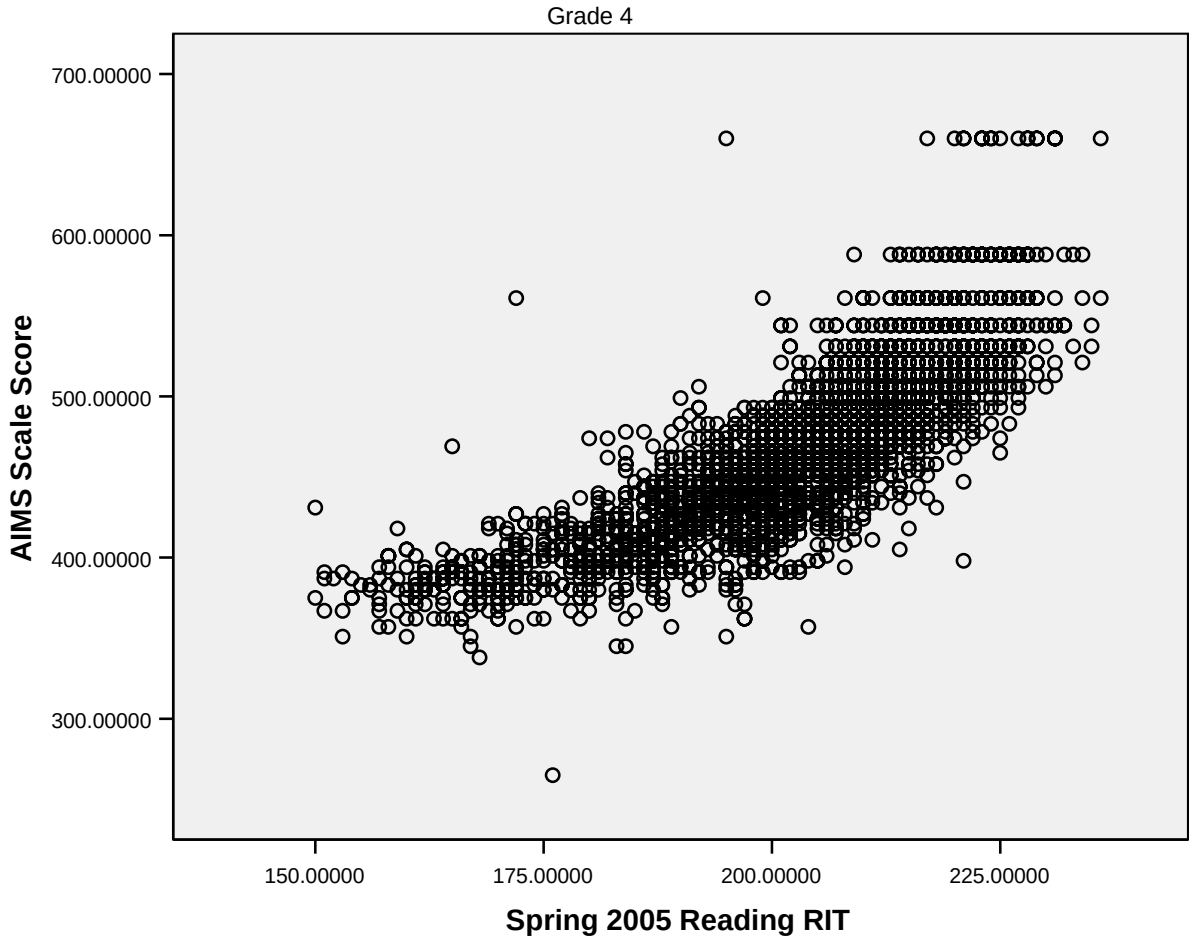
AIMS Reading							
		Spring 2005					
Grade	N*	Mean	Median	SD			
3	3146	447.61	447	51.85			
4	2961	468.75	469	52.67			
5	2904	487.80	489	48.39			
6	2720	495.21	498	48.66			
7	2686	507.21	506	52.05			
8	2660	513.71	512	51.76			
NWEA Reading							
	Spring 2005			Fall 2004			
Grade	Mean	Median	SD	N	Mean	Median	SD
3	195.34	198	16.39	2902	185.98	188	16.85
4	202.81	205	15.42	2744	195.87	199	15.86
5	208.70	211	15.41	2724	202.49	205	15.79
6	209.82	215	15.31	2277	209.82	212	14.90
7	214.83	217	16.67	2194	213.05	215	15.83
8	219.08	221	16.13	2213	217.20	219	16.17
AIMS Mathematics							
		Spring 2005					
Grade	N	Mean	Median	SD			
3	3148	446.46	445	49.44			
4	2959	477.31	478	50.49			
5	2899	506.73	506	56.37			
6	2636	519.72	515	57.36			
7	2671	538.53	538	54.26			
8	2654	548.54	548	56.67			
NWEA Mathematics							
	Spring 2005			Fall 204			
Grade	Mean	Median	SD	N	Mean	Median	SD
3	201.85	203	13.31	2906	190.25	191	12.79
4	210.50	210	14.03	2741	201.08	202	12.95
5	219.12	221	15.25	2719	209.83	211	14.32
6	224.50	226	15.74	2204	218.30	219	14.47
7	227.46	229	17.22	2185	223.00	225	16.17
8	231.94	234	17.78	2218	227.53	229	17.01

Table 2 – Inter-test Correlations for AIMS and NWEA assessments by Subject

	Reading			Mathematics		
Grade	AIMS - Spring RIT	AIMS - Fall RIT	Spring RIT-Fall RIT	AIMS - Spring RIT	AIMS - Fall RIT	Spring RIT-Fall RIT
3	.85	.83	.85	.84	.82	.83
4	.82	.80	.84	.85	.81	.85
5	.83	.82	.87	.86	.83	.87
6	.82	.81	.85	.86	.83	.88
7	.81	.79	.85	.86	.84	.89
8	.79	.78	.86	.88	.86	.89

In general, scatterplots showed that relationships between NWEA and AIMS scores were curvilinear with some evidence of floor and ceiling effects for the two ends of the scales. This would suggest that the NWEA assessment may measure the very high and low performing students with greater precision than the state test. This isn't unexpected because state assessments are typically designed to generate estimates of performance using the grade level standards and content, thus some of this effect may be a product of the limitations inherent in a grade level test's ability to deliver items that accurately measure students in the extremes of the performance range. Figure 1 shows an example that is illustrates both the strength of the linear correlation and the issue of dispersion.

Figure 1 – Grade 4 Reading AIMS Reading score plotted against Spring Reading RIT score



Linking AIMS performance level cut scores to the RIT scale

The primary purpose of this study was to generate new estimates of the RIT scale scores that most closely correspond to the cut scores for different performance levels on the AIMS. This information allows schools to identify students who may need additional support to reach state standards. It can also help

schools identify students who are performing well enough that they are ready to tackle work beyond what the state standards require.

Our alignment studies employ three methods to estimate cut scores, linear regression, second order regression, and a Rasch status on standards (Rasch SOS) method that estimates cut scores using a design based in item-response theory.

Tables 3 and 4 show several estimations of the spring and prior fall RIT scores that correspond to the cut scores for the various performance levels on the AIMS scales. As a rule the three methodologies came to similar estimates of cut scores for each of the performance levels with almost all estimates falling within a three point range.

Table 3 – Estimated points on the RIT scale equating to the minimum scores (rounded) for performance levels on the AIMS based on SPRING testing

Grade 3												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<169	169	189	221	<168	168	191	218`	<168	168	190	219
Mathematics	<183	183	194	217	<180	180	195	215	<179	179	194	215
Grade 4												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<179	179	196	228	<178	178	198	222`	<179	179	198	227
Mathematics	<190	190	201	225	<189	189	202	225	<191	191	201	224
Grade 5												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<184	184	201	236	<184	184	204	231	<184	184	203	231
Mathematics	<199	199	210	233	<198	198	211	232	<197	197	210	232
Grade 6												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<190	190	207	242	<188	188	209	236	<189	189	208	239
Mathematics	<206	206	217	242	<207	207	219	241	<206	206	218	241
Grade 7												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<190	190	208	247	<189	189	210	239	<187	187	210	240
Mathematics	<208	208	220	250	<208	208	222	247	<207	207	221	249
Grade 8												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<195	195	213	254	<196	196	216	243	<193	193	215	247
Mathematics	<217	217	228	259	<218	218	230	254	<217	217	230	255

Table 3 – Estimated points on the RIT scale equating to the minimum scores (rounded) for performance levels on the AIMS based on prior FALL testing

Grade 3												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<158	158	179	212	<157	157	180	211	<157	157	178	212
Mathematics	<171	171	182	204	<169	169	182	204	<168	168	181	203
Grade 4												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<170	170	188	222	<168	168	191	217	<170	170	190	221
Mathematics	<180	180	191	215	<180	180	192	215	<181	181	192	214
Grade 5												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<176	176	194	230	<174	174	196	224	<174	174	195	225
Mathematics	<189	189	200	223	<188	188	201	222	<187	187	200	222
Grade 6												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<184	184	202	238	<182	182	204	232	<182	182	203	235
Mathematics	<200	200	210	234	<199	199	211	233	<199	199	211	233
Grade 7												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<187	187	205	243	<186	186	208	236	<184	184	207	238
Mathematics	<203	203	214	244	<202	202	216	241	<202	202	215	243
Grade 8												
	Linear				Second-order regression				Rasch SOS			
	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds	Far Below	App	Meets	Exceeds
Reading	<192	192	210	251	<193	193	214	242	<190	190	212	247
Mathematics	<212	212	223	253	<214	214	226	250	<212	212	224	250

Establishing RIT score estimates for AIMS performance levels.

Once the cut scores were estimated from the three methods, we evaluated each set of possible cut scores to determine how accurately it predicted students' actual performance on the corresponding AIMS assessment. The most accurate method of prediction was generally used to derive the best estimate of RIT cut scores that equate to the different AIMS performance levels.

The following methods were used to establish the most accurate method for each performance level:

- **Falls Far Below and Approaches**. We selected the RIT cut score that correctly identified the largest proportion of students who performed at the **Falls Far Below** level on the AIMS assessment.

- **Meets.** We calculated a *prediction index* statistic for the proposed **Meets** cut score. This is calculated as 1 – (correct predictions/type I errors). Correct predictions occur when the cut score that is established accurately predicts passage on the state assessment for a student. Type 1 errors occur when students who scored at or above the cut score do not pass the state test. A test with a high prediction index statistic typically reflects both a high rate of accuracy and a low rate of Type I errors. We generally selected the method that produced the highest prediction index number.

- **Exceeds.** We selected the method that correctly identified the largest proportion of students who scored in the **Exceeds** category on the AIMS.

Tables 4 through 7 show the recommended RIT cut scores for each of the AIMS performance levels.

Table 4 – Recommended SPRING RIT cut scores for AIMS performance levels – Reading

	Far Below			Approaches			Meets			Exceeds		
Grade	Score	Method	% of far below students found	Score			Score	Method	Prediction Index	Score	Method	% of exceeds students found
3	<169	L	57.7%	169			191	S	.909	218	S	35.5%
4	<179	L,R	55.2%	179			198	R	.882	222	S	40.8%
5	<184	L,R	52.4%	184			204	S	.908	231	S,R	27.0%
6	<190	L	47.1%	190			209	S	.898	236	S	28.2%
7	<190	L	51.8%	190			210	S	.882	239	S	35.3%
8	<196	S	44.4%	196			216	S	.873	243	S	27.5%

(L= Linear Regression, S=Second Order Regression, R=Rasch SOS method)

Table 5 – Recommended PRIOR FALL RIT cut scores for AIMS performance levels – Reading

	Far Below			Approaches			Meets			Exceeds		
Grade	Score	Method	% of far below students found	Score			Score	Method	Prediction Index	Score	Method	% of exceeds students found
3	<158	L	50.9%	158			180	S	.907	211	S	25.9%
4	<170	L	50.3%	170			191	S	.897	217	S	38.8%
5	<176	L	54.4%	176			196	S	.878	224	S	28.4%
6	<184	L	48.2%	184			204	S	.897	232	S	31.1%
7	<187	L	43.8%	187			208	S	.877	236	S	31.4%
8	<193	S	44.2%	193			214	S	.877	242	S	20.8%

(L= Linear Regression, S=Second Order Regression, R=Rasch SOS method)

Table 6 – Recommended SPRING RIT cut scores for AIMS performance levels – Mathematics

	Far Below			Approaches			Meets			Exceeds		
Grade	Score	Method	% of far below students found	Score			Score	Method	Prediction Index	Score	Method	% of exceeds students found
3	<183	L	52.2%	183			195	S	.898	215	S,R	67.2%
4	<191	R	51.5%	191			202	S	.902	224	R	68.5%
5	<199	L	59.4%	199			211	S	.924	232	R	71.8%
6	<206	L	59.3%	206			219	S	.909	241	S	65.2%
7	<208	L	62.9%	208			222	S	.912	247	S	66.4%
8	<217	L	71.1%	217			230	S,R	.918	254	S	60.8%

(L= Linear Regression, S=Second Order Regression, R=Rasch SOS method)

Table 7 – Recommended PRIOR FALL RIT cut scores for AIMS performance levels – Mathematics

	Far Below			Approaches			Meets			Exceeds		
Grade	Score	Method	% of far below students found	Score			Score	Method	Prediction Index	Score	Method	% of exceeds students found
3	<171	L	42.8%	171			182	S	.885	203	R	63.2%
4	<181	R	44.1%	181			191	L	.883	214	R	52.5%
5	<189	L	48.2%	189			200	L	.909	222	S,R	67.3%
6	<200	L	58.9%	200			211	S	.895	233	S	66.3%
7	<203	L	56.9%	203			216	S	.903	241	S	66.9%
8	<212	L,R	64.2%	212			226	S	.917	250	S,R	53.8%

(L= Linear Regression, S=Second Order Regression, R=Rasch SOS method)

We evaluate the relative accuracy of state alignment study results by comparing the prediction index statistics generated by these studies for their accuracy in assessing proficiency status and performance level. The results show that the Arizona studies were in the lower third of all studies relative to accuracy of pass-fail prediction. In terms of performance level prediction, the results in reading were in the upper half of all studies while the results in mathematics were in the lower half.

Table 8 – Prediction Indices (Based on Proficiency Status) for Previous NWEA State Alignment Studies

State	Reading	State	Language	State	Math
Texas	.967*	Texas	.968*	Texas	.969*
Minnesota	.944*	South Carolina Exit	.938*	Wyoming	.961
South Carolina Exit	.940*	California	.913*	Colorado '01	.957
Pennsylvania	.935*	Indiana '01	.907*	Illinois	.946*
Wyoming	.931	Colorado '03	.903*	Colorado '03	.943*
Colorado '03	.931*	Indiana '03	.894*	South Carolina '03	.943*
Illinois	.928*	Indiana '05	.891	Minnesota	.936*
California	.925*	South Carolina '04	.889*	South Carolina Exit	.933*
Arizona '03	.912*	Arizona '03	.874*	Pennsylvania	.926*
Colorado '01	.910*			Washington '99	.920
Nevada	.902*			Arizona '03	.919*
South Carolina '03	.902*			South Carolina '04	.914*
Indiana '01	.902*			Washington '04	.912*
Indiana '03	.900*			Arizona '05	.910
Washington '99	.893			California	.910*
Indiana '05	.892			Indiana '05	.906
Arizona '05	.891			Indiana '01	.899*
Washington '04	.886*			Nevada	.866*
South Carolina '04	.884*			Indiana '03	.860*

Table 9 – Prediction index scores by performance level assignment for previous NWEA state alignment Studies

State	Reading	State	Math
Texas	.868	Texas	.900
Indiana '05	.867	Illinois	.888*
Indiana '03	.860	Indiana '05	.863
Colorado	.840	Colorado	.808
Illinois	.804*	Indiana '03	.804*
Arizona '05	.781	Pennsylvania	.769*
Nevada	.776*	South Carolina '03	.764*
Pennsylvania	.770*	Nevada	.742*
South Carolina '03	.757*	South Carolina '04	.741*
Arizona '03	.756*	Arizona '05	.730
South Carolina '04	.717*	Arizona '03	.726
Washington '04	.667	Washington '04	.721
South Carolina Exit	.649*	South Carolina Exit	.705*
Minnesota	.627*	Minnesota	.611*
California	.600*	California	.565*

Using RIT scores to estimate student probability of achieving passing performance on the AIMS

Although the predicted RIT cut scores can help teachers and students establish targets for NWEA assessments that can help assure success on the state test, teachers should be aware that students performing near the proficient cut score on the RIT scale have only about a 50% probability of passing the AIMS. The information in Tables 10 through 13 report more precise data related to students' probabilities of achieving proficiency.

These tables show the proportion of students at each 5 point RIT level who earned scores at or above the *proficient* level on their respective AIMS assessment both when the NWEA test was administered in the same season as the state test (spring) and also when the NWEA test was administered during the prior fall. Using Table 10 as an example, we would find that about 26% of the Grade 4 students who achieved a reading RIT score between 190 and 194 went on to achieve a score of **Meets** on the AIMS assessment. A reading teacher would know that only about one in four of students performing in this range in spring is likely to achieve a proficient score on the AIMS unless they work harder, receive more focused instruction, or have access to additional resources.

On the other hand, about 83% of students who scored between RITs of 205 and 209 achieved **Meets** on the Arizona assessment. Teachers should feel free to focus their efforts with these students on content and skills that go beyond the minimum expectations for performance.

Figures 3 through 8 are graphic depictions of the data in the tables.

Table 10 – Proportion of students passing the AIMS Reading based on SPRING RIT reading score

RIT	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
145	0.00%					
150	2.70%					
155	1.72%					
160	3.33%					
165	1.05%					
170	5.50%		0.00%			
175	11.33%	0.00%	4.65%			
180	11.86%	6.90%	3.08%		0.00%	
185	31.66%	7.36%	2.68%	0.00%	1.75%	
190	52.78%	25.97%	5.26%	3.81%	4.35%	
195	79.85%	38.18%	19.90%	12.50%	8.55%	1.28%
200	90.57%	64.47%	48.12%	21.70%	21.51%	9.52%
205	96.72%	82.89%	67.80%	43.97%	31.84%	14.94%
210	98.25%	95.28%	86.92%	65.36%	54.03%	37.10%
215	100.00%	98.54%	96.35%	84.06%	72.98%	50.00%
220		99.06%	99.71%	95.00%	87.39%	75.07%
225		100.00%	100.00%	98.66%	97.50%	89.37%
230				99.49%	98.73%	97.49%
235				100.00%	100.00%	98.39%
240						100.00%

Table 11 - Proportion of students passing the AIMS Reading based on PRIOR Fall RIT reading score

RIT	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
140	0.00%					
145	2.27%					
150	1.16%					
155	5.74%	0.00%	0.00%			
160	8.77%	1.82%	2.78%			
165	14.29%	6.94%	2.33%			
170	25.00%	6.74%	3.85%			
175	40.99%	10.53%	8.62%	0.00%	0.00%	
180	60.28%	20.71%	10.99%	6.25%	2.86%	
185	78.03%	36.55%	16.33%	7.94%	11.11%	0.00%
190	90.72%	53.93%	36.16%	12.38%	7.95%	2.13%
195	97.10%	76.81%	51.98%	19.30%	10.69%	1.43%
200	97.88%	88.94%	74.64%	50.00%	34.48%	13.79%
205	99.08%	94.82%	88.97%	64.89%	43.89%	30.99%
210	100.00%	98.38%	95.93%	83.93%	68.00%	44.44%
215		99.38%	100.00%	94.29%	82.22%	66.55%
220		100.00%		97.60%	93.86%	82.54%
225				100.00%	97.50%	93.91%
230					99.35%	99.16%
235					100.00%	99.32%
240						100.00%

Table 12– Proportion of students passing the AIMS mathematics test based on SPRING RIT Mathematics Score

RIT	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
150						
155	0.00%					
160	5.77%					
165	9.20%	0.00%				
170	18.78%	7.32%	0.00%			
175	27.43%	4.65%	5.13%			
180	49.44%	10.08%	4.05%			
185	73.87%	20.99%	7.29%			
190	86.82%	45.52%	13.79%	0.00%		0.00%
195	95.49%	74.00%	30.41%	7.21%	0.00%	3.64%
200	100.00%	86.41%	56.09%	12.82%	7.52%	2.35%
205		95.67%	77.31%	26.42%	20.28%	4.76%
210		98.78%	90.03%	49.80%	37.21%	9.87%
215		99.45%	96.94%	74.73%	54.13%	18.08%
220		99.12%	98.94%	90.54%	73.75%	40.74%
225		100.00%	100.00%	97.52%	89.57%	58.05%
230				98.80%	97.19%	81.65%
235				100.00%	99.55%	96.06%
240					100.00%	97.88%
245						99.47%
250						100.00%

Table 13 Proportion of students passing the AIMS mathematics test based on PRIOR FALL Mathematics RIT score

RIT	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
165	0.00%	0.00%				
170	6.52%	8.33%				
175	6.93%	2.70%	0.00%			
180	13.19%	4.48%	2.86%			
185	11.39%	4.12%	1.92%			
190	35.59%	12.33%	2.63%			
195	59.39%	23.73%	4.48%		0.00%	
200	82.07%	52.87%	12.79%	0.00%	6.19%	
205	97.55%	72.09%	29.80%	9.83%	4.32%	0.00%
210	99.47%	91.22%	55.48%	15.70%	10.76%	2.13%
215	100.00%	96.46%	81.63%	41.83%	31.98%	9.04%
220		99.66%	93.70%	65.79%	48.41%	14.55%
225		100.00%	97.09%	84.26%	71.96%	29.60%
230			99.68%	95.43%	88.95%	59.57%
235			100.00%	98.83%	96.89%	81.38%
240				99.46%	100.00%	94.86%
245				99.27%	99.48%	99.27%
250				100.00%	100.00%	100.00%

Figure 2 – Percent of Students Passing AIMS reading by Spring Reading RIT Performance Range

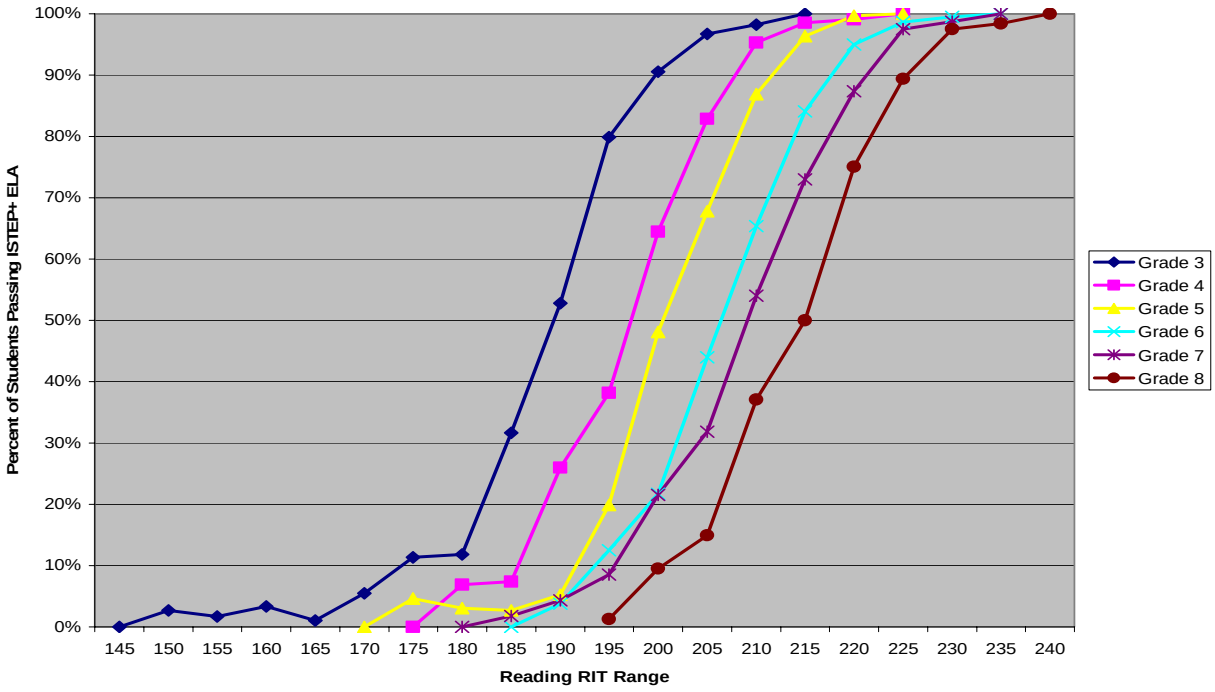


Figure 3 – Percent of Students Passing AIMS reading by PRIOR FALL Reading RIT Performance Range

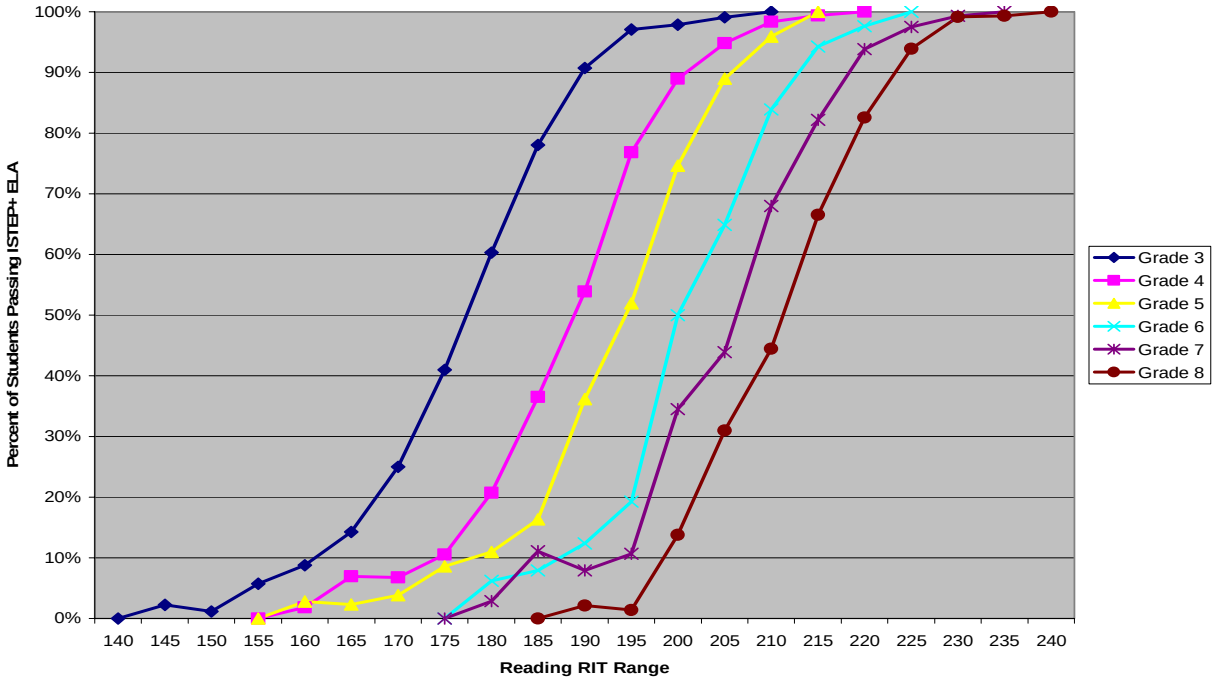


Figure 4 – Percent of Students Passing AIMS Mathematics by SPRING Mathematics RIT Performance Range

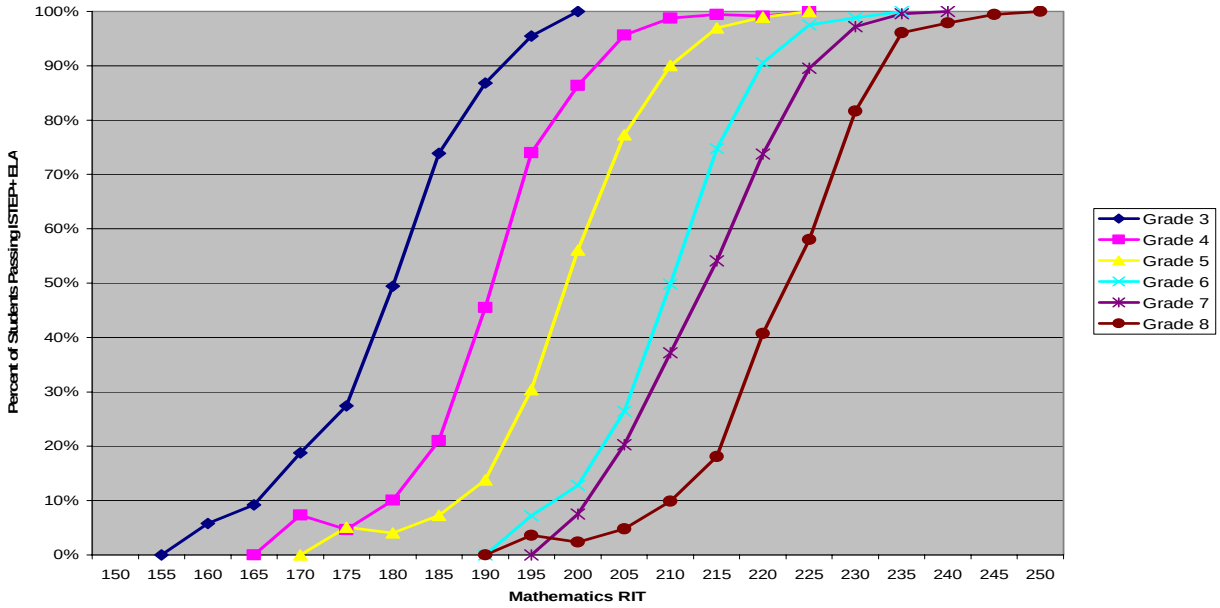
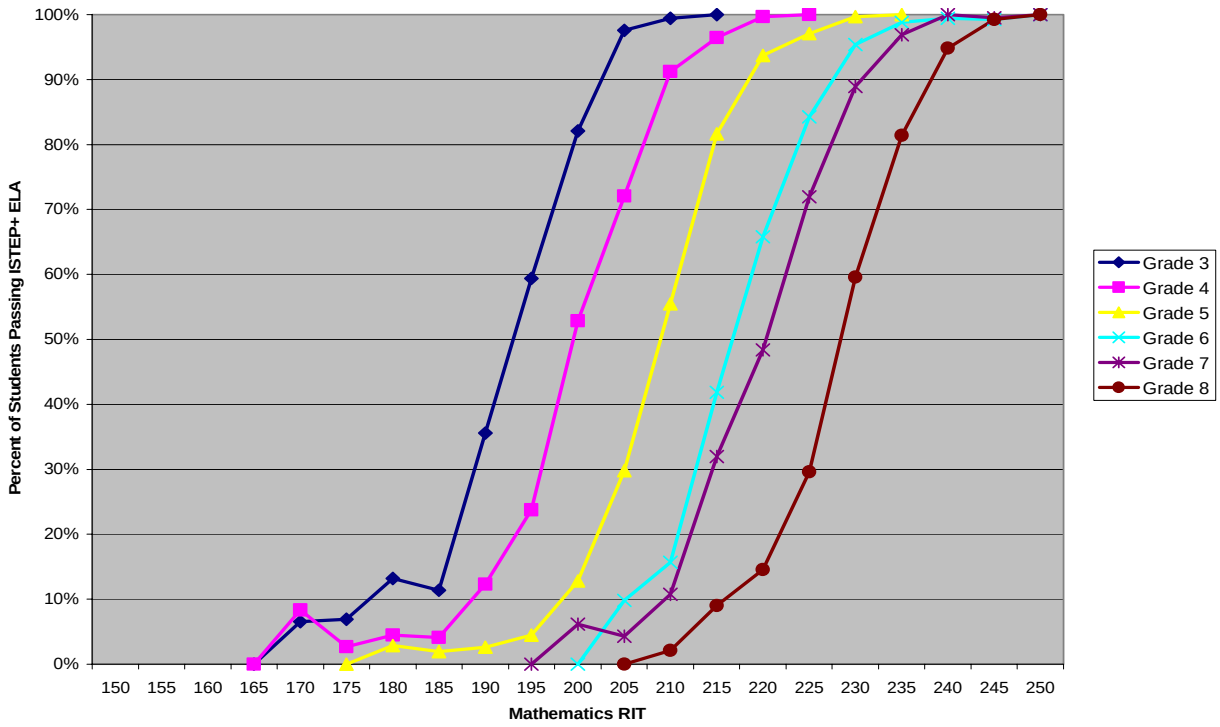


Figure 5 – Percent of Students Passing AIMS Matheamtics by PRIOR FALL Mathematics RIT Performance Range



Comparing changes in the estimated AIMS standards relative to the prior alignment study

Scale Stability and Cut Score Changes

It is impossible to accurately measure improvement without maintaining a stable scale. Imagine that Sid is working on his golf game and that he uses the 250 yard marker at the local driving range to estimate his driving distance. He watches 40% of his drives roll beyond this marker in his first practice session. Sid does some weight work and takes a couple of lessons from his pro and returns to the driving range two weeks later. Now 60% of Sid's drives roll beyond the 250 marker. He naturally assumes that practice has led to improvement.

Suppose, however, that the range manager had moved the tee boxes forward ten yards so that golfers would hit off of fresh grass. If that happened, then we don't know whether Sid's improvement was a product of his hard work or a product of a change in the scale. In other words, the 250 yard marker represents a different distance today than it did two weeks ago.

Similarly, it's impossible to measure improvement on an academic test without maintaining a very stable scale. Even small changes in a test's difficulty relative to its predecessors can have a noticeable effect on proficiency rates that is independent of instruction. If a test is slightly easier than its predecessor's for example, proficiency rates may improve (just like Sid's driving distance seemed to improve) without an actual improvement in learning having occurred.

Important modifications were made on the 2005 version of the AIMS assessment. The test design, scale design, and cut scores all changed. As a result of these changes, statewide pass rates on the AIMS increased substantially this year over the prior year. In mathematics, for example, the statewide pass rate across grades improved from 31% to 63%. In reading, the statistic improved from 49% to 69%.

These changes resulted in estimated RIT cut score estimates that were significantly lower than those estimated from the 2003 study, especially in the middle and upper grades (see Table 14). The grade 5 reading and mathematics **meets** cut scores, fore example, were 6 and 10 RIT points lower respectively than the scores estimated from the 2003 study. The grade 8 reading and mathematics cut scores were 8 and 18 points lower than the 2003 estimates.

Applying these differences to the 2005 NWEA norms, we find that large numbers of additional students will achieve passing scores without necessarily improving their performance. Using the grade 5 estimates in reading as an example, the 6 RIT difference in cut scores would result in 17% more students reaching the standard in 2005 than would have reached the standard in 2003 in a district with a performance distribution that reflects our current norms.

The change that we are seeing in Arizona is consistent with changes that we have seen recently in other states in which we have conducted multiple alignment studies. We have recently completed follow-up alignment studies in Washington, Indiana, and South Carolina. While cut scores did not change in a single consistent direction in South Carolina, we found substantially lower cut scores for both Washington and Indiana in these follow-up studies.

Table 14 – Estimated RIT cut scores for the Proficient level of performance on the AIMS 2001-2005*

	Reading		Mathematics	
	2003	2005	2003	2005
Grade 3	190 (26)	191(28)	199 (39)	195 (27)
Grade 4		198 (27)		202 (26)
Grade 5	210 (44)	204 (27)	220(54)	211 (31)
Grade 6		209 (30)		219 (38)
Grade 7		210 (25)		222 (34)
Grade 8	224 (53)	216 (31)	248 (78)	230 (40)

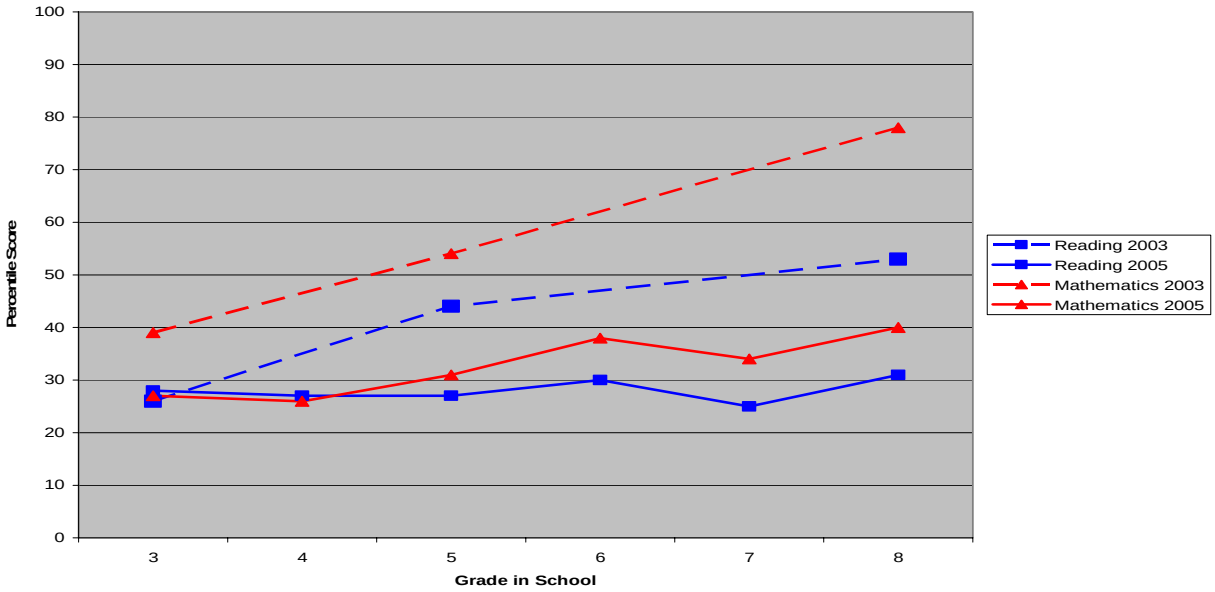
*NWEA percentile score (based on 2005 norms study) is in parentheses

Calibration

It is also desirable for proficiency cut scores to calibrate across grades. By this we mean that the proficiency standard for performance in one subject and great, say grade 3 mathematics for example, should be no easier or more difficult than the standard at other grades. There is an argument to be made as well for consistency across subjects. If the mathematics standard is going to be set at a level that makes it more difficult for students to achieve than a reading standard, people should be aware of that fact when establishing the standard and communicate their rationale for making such a decision.

Our 2003 study found that the AIMS Reading and Mathematics tests did not calibrate very well across grades (see Figure 6). In mathematics, for example, a student performing near the 30th percentile could achieve **meets** performance on AIMS in grade 3, while an 8th grade student would need to perform near the 80th percentile to achieve that designation in grade 8. Our 2005 study found much closer calibration of the standards across grades in both reading and math (the flatter trajectory of the lines in Figure 5 provide the evidence). In mathematics for example, the 2005 standard is set at about the 30th percentile for grade 3 and at the 40th percentile for grade 8.

Figure 6 –2003 and 2005 estimated RIT scores for Meets performance in Reading and Mathematics at each grade



The standards also calibrated a bit more closely across subjects. The 2003 mathematics standards were far more challenging than the reading standards. While the 2005 mathematics standards are still more difficult than the reading standards (relative to the RIT scale), the differences are far smaller.

Comparing the AIMS standards relative to those in place in other states

Northwest Evaluation Association tests have been aligned with the cut scores state assessments in 23 states. To get an estimate of the difficulty of the AIMS in relation to other state tests, we evaluated the standard defined as the NCLB passing score and compared it to the cut score representing the same standard in these other states. You can view the results of this analysis at the following web location:

http://www.nwea.org/research/national.asp

In general, we believe standards should be judged on how well they align with the purposes the community has set for establishing performance expectations, not purely on how high or low the "bar" is set. If the purpose of a performance expectation is to assure that all students passing a standard will be ready to attend four year university, then the standard will need to be relatively high. On the other hand, if the purpose of a performance expectation is to assure that all students passing it graduate with the basic reading and math skills needed for entry level employment, the standard will be lower. It is clear from the evidence we've collected so far that proficiency is not yet a concept with a shared definition, because performance standards vary greatly from state to state. It would be fair to say, however, that most states that we have studied who have set standards since implementation of No Child Left Behind has begun have tended to establish standards near or below the 50th percentile on our norms. It would also be fair to say that states that have purposefully changed their performance level cut scores have moved to make their standards easier.

Summary and Conclusions

This study investigated the relationship between the scales used for the AIMS assessments and the RIT scales used to report performance on Northwest Evaluation Association tests. The study estimated the changes in reading and mathematics RIT score equivalents for the AIMS performance levels in those subjects. Test records for more than 20,000 students were included in this study.

Three methods generated an estimate of RIT cut scores that could be used to project AIMS performance levels. Accuracy of predicting AIMS passing performance was well above 80% for all grades and subjects studied when using the best methodology.

Readers should exercise some caution about generalizing these results to their own settings. Curricular or instructional differences unique to your districts may influence the accuracy with which the estimated cut scores reflect actual performance in your setting. With this limitation in mind, we would encourage educators to use this data as one tool to inform standards-based decisions.

The information gathered in this study came from measures employing the NWEA RIT Scale. Because all of the research that we have to date indicates that scores generated from computer-based tests and Achievement Level Test (ALT) scores are virtually interchangeable, readers should feel comfortable applying the results of this study in any setting that uses the RIT scale.

We hope that data from this study provides useful information to help Arizona educators use NWEA assessments to better inform, plan and deliver student instruction. Good information, when matched with the professionalism and commitment of our Arizona colleagues, will assure that every student has the opportunity to reach their aspirations.

References

Cronin J, (2003). Aligning the NWEA RIT Scale with *the Arizona Instrument to Measure Standards (AIMS)* Lake Oswego, OR. Northwest Evaluation Association.

Kingsbury, G., Olson, A., Cronin, J., Hauser, C., Houser, R. (2003). *The State of State Standards: Research Investigating Proficiency Levels in Fourteen States.* Lake Oswego, OR: Northwest Evaluation Association.