A Study of the Ongoing Alignment of the NWEA RIT Scale with assessments from the Montana Comprehensive Assessment System (MontCAS)

John Cronin, Ph.D

January 2005

NWEA
Northwest Evaluation Association

Partnering to help all kids learn

A Study of the Ongoing Alignment of the NWEA RIT Scale with assessments from the Montana Comprehensive Assessment System (MontCAS)

John Cronin, Ph.D.

January, 2005

Each year, Montana students participate in testing as part of the state's assessment program. This past spring, students in grades 4, 8, and 10 took Montana Comprehensive Assessment System (MontCAS) tests in reading 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), and to inform schools and school districts of their performance.   The Montana Office of Public Instruction has developed scales that are used to assign students to one of four performance levels on these tests.

Many students who attend school in Montana also take tests developed in cooperation with the Northwest Evaluation Association (NWEA). The content of these tests are aligned with the Montana 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. Two years ago, NWEA reported alignment of the RIT scale with the Montana proficiency cut scores on the Iowa Test of Basic Skills (Cronin and Kingsbury, 2003). With the implementation of the new MontCAS assessment, we undertook a study to evaluate the relative accuracy with which the NWEA assessments continued to predict MontCAS results. The primary questions addressed in this study are:

- What RIT scores correspond to various performance levels on the MontCAS tests?

- How do these RIT scores differ from the prior estimate of performance levels based on the Iowa Test of Basic Skills?

- How well can performance on the Montana assessments be predicted from RIT scores when NWEA assessments are administered in the same time frame?

Method

Our study included over 4,000 test records from students enrolled in 6 Montana school systems. Student records were included when a student had both a valid NWEA scale score and a valid MontCAS score in the equivalent subject. Table 1 on the following page shows the number of records included for each subject.

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, for more detail about the methods we use to conduct scale alignment studies.

Results

Descriptive Statistics

Table 1 reviews descriptive statistics for the MontCAS and NWEA assessments. The median RIT scores for this sample in reading and mathematics are slightly above the median for the NWEA norm population.

Normal distributions around a nationally-normed mean are desirable but not necessarily essential when conducting alignment studies. It is more important that the sample provide reasonable numbers of students who perform at all levels on the test scales than normal distribution so 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. In this case we had excellent representation of students who performed at all performance levels.

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

MontCAS Reading			
Grade	4	8	
N	1583	1376	
Mean	258.84	260.13	
Median	264	267	
Std. Dev.	28.73	32.91	
NWEA Reading			
Grade	4	8	
N	1583	1376	
Mean	206.96	227.61	
Median	208	229	
Std. Dev.	13.68	12.86	
MontCAS mathematics			
Grade	4	8	10
N	1616	1392	1229
Mean	243.90	266.30	265.89
Median	243	267	265
Std. Dev.	32.17	26.76	24.14
NWEA Mathematics			
Grade	4	8	10
N	1616	1392	1392
Mean	210.03	238.13	254.88
Median	211	241	255
Std. Dev.	12.14	15.88	14.56

Pearson correlations

Table 2 shows the results of this analysis for each grade. Concurrent validity was tested by examining same subject Pearson correlations between the NWEA and MontCAS assessments. Same subject correlations were high. The coefficients ranged between .75 to .84, numbers that suggest the tests were generally measuring the same constructs. Discriminant validity was tested by examining same subject Pearson correlations next to correlations for the alternate subject (math against reading). The same subject correlations were higher than correlations against the alternate subject in all subjects and grades tested.

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

	MontCAS		NWEA	
Grade 4				
	Reading	Mathematics	Reading	Mathematics
MontCAS Reading	1	.73	**.82**	.68
MontCAS Mathematics	1583	1	.73	**.75**
NWEA Reading	**1583**	1583	1	
NWEA Mathematics	1616	**1616**		1
Grade 8				
	Reading	Mathematics	Reading	Mathematics
MontCAS Reading	1	.70	**.79**	.69
MontCAS Mathematics	1376	1	.70	**.84**
NWEA Reading	**1376**	1376	1	
NWEA Mathematics	1392	**1392**		1
Grade 10				
	Reading	Mathematics	Reading	Mathematics
MontCAS reading	1	.66		.62
MontCAS Mathematics	1229	1		**.80**
NWEA Reading			1	
NWEA Mathematics	1229	**1229**		1

* shaded cells show counts

Based on a review of scatterplots, we found linear relationships between NWEA and MontCAS test scores, although some evidence of floor and ceiling effect were evident. Figure 1 shows an example that illustrates floor effect using a scatterplot of the Grade 4 reading results. The figure shows a fairly large cluster of students at the lowest scale score. These same students have RIT scores that range from 150 to nearly 210 RIT. Lack of student motivation may be part of the explanation for this, as might a lack of measurement sensitivity in the lowest part of the scale due to the grade level test design. Figure 2 shows an example of ceiling effect that was evident at grade 10. In this case a fairly large cluster of students performed at the highest awarded score on the MontCAS. Because the state test must be written to test only standards that all students should have encountered by grade 10, very high performing students may top out this kind of assessment. NWEA's high school assessment, which are broken down by subject at the high school level, do not have this limitation and thus the scale has more range to measure student performance at the high end.

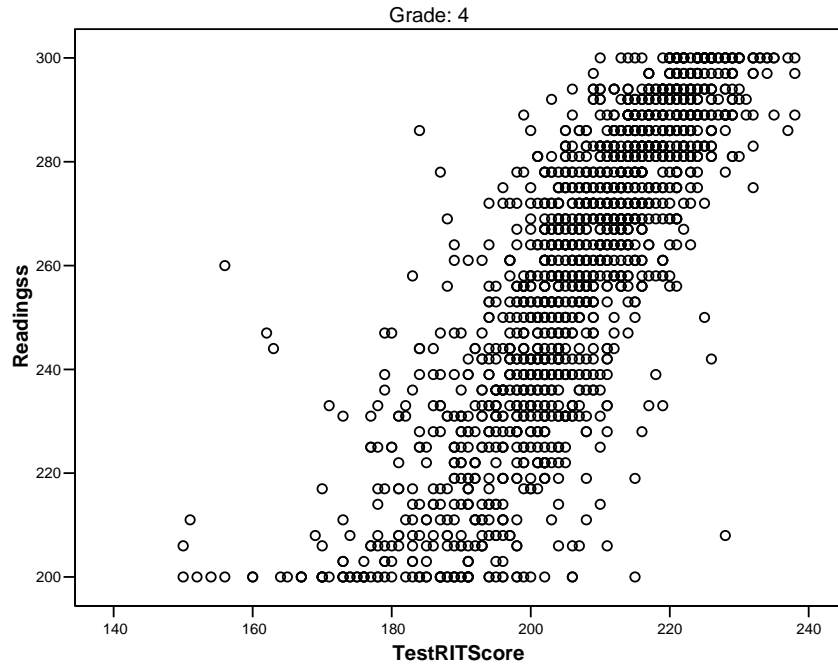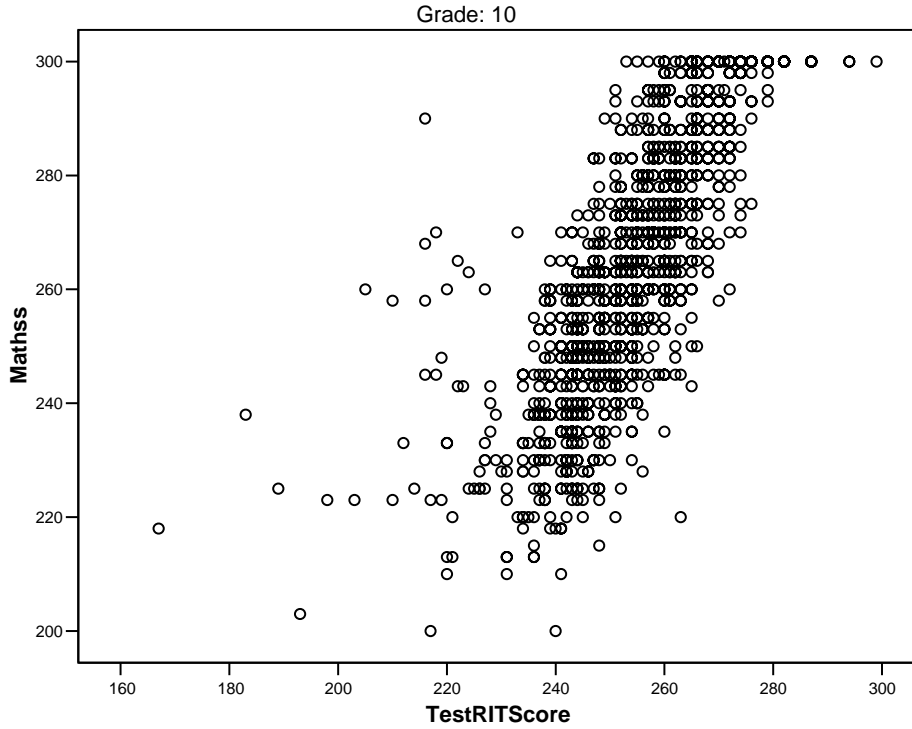# Figure 1 – Grade 7 Reading MontCAS score plotted against Reading RIT score

Grade: 4

Figure 2 – Grade 10 Math MontCAS score plotted against Math RIT score

Grade: 10

Linking MontCAS 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 MontCAS. 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.

Table 3 shows several estimations of the Spring 2003 RIT score that correspond to the cut scores for the various performance levels on the MontCAS scales. As a rule the three methodologies came to similar estimates of cut scores for each of the performance levels, although the Rasch SOS methodology did produce somewhat higher estimates of the RIT score required to meet the basic standard at some grades.

Table 3 – Estimated points on the RIT scale equating to the minimum scores (rounded) for performance levels on the MontCAS

	Grade 4											
	Linear Regression				Second-order Regression				Rasch Status-on-Standard			
	N	NP	P	A	N	NP	P	A	N	NP	P	A
Reading	<188	188	202	220	<187	187	202	220	<193	193	203	216
Mathematics	<202	202	213	230	<201	201	214	227	<205	205	212	223
	Grade 8											
	Linear Regression				Second-order Regression				Rasch Status-on-Standard			
	N	NP	P	A	N	NP	P	A	N	NP	P	A
Reading	<211	211	223	238	<210	210	224	238	<217	217	224	234
Mathematics	<209	209	227	258	<208	208	231	255	<208	208	232	250
	Grade 10											
	Linear Regression				Second-order Regression				Rasch Status-on-Standard			
	N	NP	P	A	N	NP	P	A	N	NP	P	A
Mathematics	<226	226	244	270	<214	214	242	269	<228	228	246	265

Establishing RIT score estimates for MontCAS 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 MontCAS assessment. The most accurate method of prediction was generally used to derive the best estimate of RIT cut scores that equate to the different MontCAS performance levels.

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

- **Novice and Nearing Proficiency.** We selected the method that correctly identified the largest portion of students who scored in the *below basic* category on MontCAS.

- **Proficient.** We calculated a *prediction index* statistic for the proposed cut score. This is calculated as 1 – (correct predictions/type I errors). 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.

- **Advanced.** We selected the method that correctly identified the largest proportion of students who scored in the *advanced* category on the MontCAS.

Tables 4 and 5 show the recommended RIT cut scores for each of the MontCAS performance levels. In general, Rasch SOS methods were most reliable for establishing predictive cut scores for the highest and lowest performance levels, while all methods were similarly effective for predicting performance at the proficient level. For years in which MontCAS is not currently administered, we offer interpolated scores for the grades based on the growth we project would be needed to maintain status in that category

For proficiency status, the rate of correct prediction was above 80% for all grades. For students at the highest and lowest performance level, the recommended cut score correctly identified between about 54% and 79% of the students who actually performed in that category.

Table 4 – Recommended RIT cut scores for MontCAS performance levels - Reading

	Novice			Nearing Proficiency	Proficient			Advanced		
Grade	Score	Method	% of students ID	Score	Score	Method	Prediction Index*	Score	Method	% of students ID
3	<179			179	192			209		
4	<193	R	63.6%	193	203	R	.913 (85%)	216	R	77.2%
5	<200			200	210			222		
6	<207			207	215			226		
7	<213			213	220			230		
8	<217	R	67.6%	217	224	S,R	.891 (84%)	234	R	73.4%

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

- percent of students correctly predicted is in parentheses.

- Projected on-track cut scores are interpolated (in blue) for years in which MontCAS is not administered

Table 5 – Recommended RIT cut scores for MontCAS performance levels – Mathematics

	Novice			Nearing Proficiency	Proficient			Advanced		
Grade	Score	Method	% of students ID	Score	Score	Method	Prediction Index*	Score	Method	% of students ID
3	<194			194	204			215		
4	<205	R	70.79%	205	214	S	.934 (82%)	223	R	57.38%
5	<206			206	219			231		
6	<207			207	223			238		
7	<208			208	228			244		
8	<209	L	50.00%	209	232	R	.921 (87%)	250	R	75.29%
9	<218			218	242			258		

10	<228	R	27.27%	228	246	R	.898 (84%)	265	R	80.56%

We evaluate the relative accuracy of state alignment studies by comparing the prediction index statistics generated by these studies for accuracy in assessing proficiency status and performance level. Table 6 summarizes the accuracy of proficiency status prediction for this study relative to other state alignment studies and Table 7 summarizes the accuracy of performance level prediction. The results show that the prediction index statistics for proficiency status in reading are near average when compared to other state studies and low in mathematics.

Table 6 – 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	South Carolina '04	.889	Minnesota	.936
California	.925	Arizona	.874	South Carolina Exit	.933
Arizona	.912			Pennsylvania	.926
Colorado '01	.910			Washington '99	.920
Montana	.903			Arizona	.919
Nevada	.902			South Carolina '04	.914
South Carolina '03	.902			Washington '04	.912
Indiana '01	.902			California	.910
Indiana '03	.900			Montana	.899
Washington '99	.893			Indiana '01	.899
Washington '04	.886			Nevada	.866
South Carolina '04	.884			Indiana '03	.860

Table 7 – Prediction index scores by performance level assignment for previous NWEA state alignment Studies

State	Reading	State	Math
Texas	.868	Texas	.900
Indiana	.860	Illinois	.888
Colorado	.840	Colorado	.808
Illinois	.804	Indiana	.804
Nevada	.776	Pennsylvania	.769
Pennsylvania	.770	South Carolina '03	.764
South Carolina '03	.757	Arizona	.726
Arizona	.756	Nevada	.742
South Carolina '04	.717	South Carolina '04	.741
Montana	.670	Washington	.721
Washington	.667	Montana	.707
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 MontCAS

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 MontCAS. The information in Tables 8 and 9 provide educators with 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 MontCAS assessment. Using reading as an example, we find that about 25% of the Grade 4 students who achieved a reading RIT score between 195 and 199 went on to achieve a proficient score on the MontCAS assessment. A reading teacher would know that only about one in four of these students is likely to achieve a proficient score on the MontCAS unless they work harder, receive more focused instruction, or have access to additional resources.

On the other hand, about 93% of students who scored between RITs of 210 and 214 achieved proficiency on the Montana 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 and 4 are graphic depictions of the data in the tables.

Table 8 – Proportion of students passing the MontCAS reading based on same spring RIT reading score

| RIT	Grade 4	Grade 8
175	0.00%	
180	5.13%	
185	8.47%	
190	9.18%	
195	24.11%	
200	50.00%	0.00%
205	75.45%	2.33%
210	92.70%	10.29%
215	97.00%	22.13%
220	100.00%	45.14%
225	97.67%	71.01%
230	100.00%	88.28%
235		97.91%
240		100.00%

Table 9 – Proportion of students passing the MontCAS mathematics based on same spring RIT mathematics score

RIT	Grade 4	Grade 8	Grade 10
180	0.00%		
185	3.45%		
190	1.37%		
195	5.43%	0.00%	
200	9.90%	5.00%	
205	22.96%	6.90%	
210	51.19%	15.00%	
215	76.75%	4.76%	
220	92.97%	17.65%	
225	94.23%	30.77%	7.14%
230	94.59%	57.14%	4.35%
235	90.00%	73.94%	23.68%
240	100.00%	91.39%	33.77%
245		98.32%	59.73%
250		99.53%	77.56%
255		100.00%	92.47%
260			94.37%
265			99.26%
270			100.00%

Figure 3 – Percent of Students Passing Reading MontCAS by RIT Performance Range

Percent of Students Passing Reading MontCAS based on RIT Range
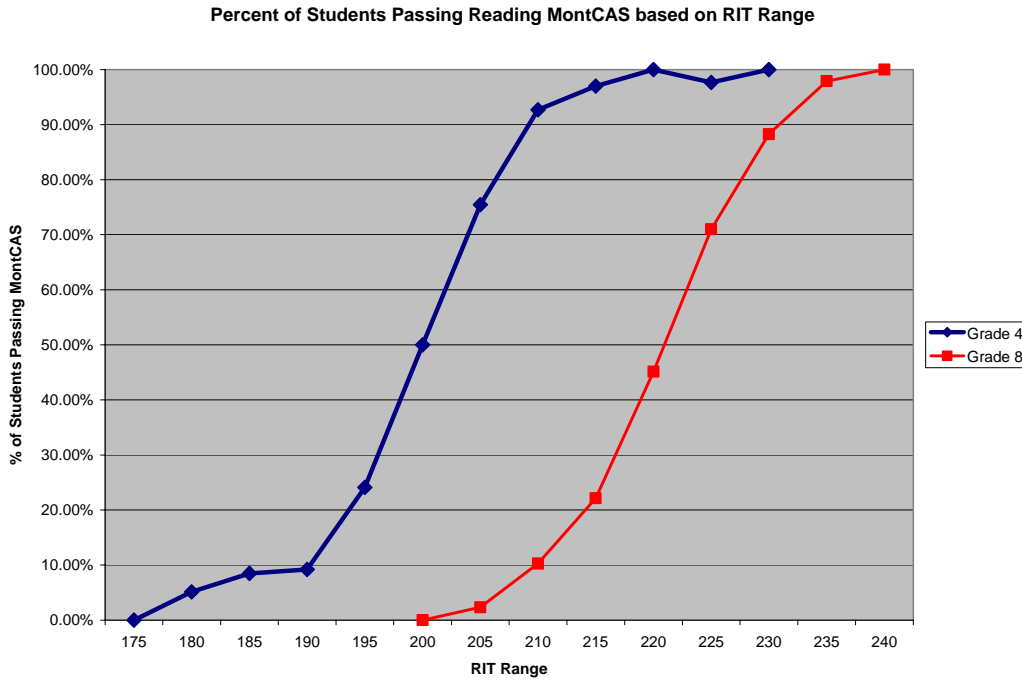
Figure 4 – Percent of Students Passing Mathematics MontCAS by RIT Performance Range

Percent of Students Passing Math MontCAS based on RIT Range
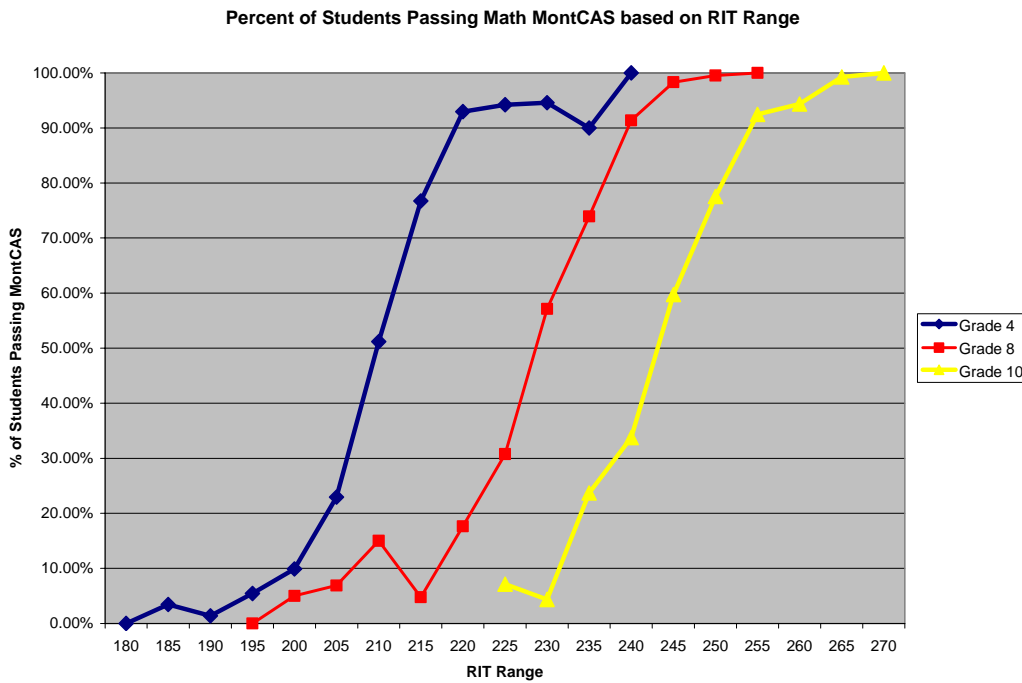
Comparing changes in the estimated MontCAS standards relative to the prior alignment study

Table 10 compares the proficient level cut scores found for the current test with estimates that NWEA derived from the standard used with the state's prior test, the Iowa Test of Basic Skills (ITBS). \

Our estimates indicate that the proficient level cut scores for the MontCAS reading are higher than those that were in place for the ITBS in reading. The differences, 7 points in grade 4 and 5 points in grade 8 represent nearly one full year of growth in those grades. The estimated mathematics cut scores for the MontCAS are higher than those for estimated from the ITBS in grades 4 and 8. The 10th grade cut score is nearly the same.

Table 10 – Estimated RIT cut scores for the Proficient level of performance on the MontCAS 1999/2004*

| Grade	Reading		Mathematics	
	ITBS	MontCAS	ITBS	MontCAS
4	196 (26)	203 (42)	205 (39)	214 (66)
8	219 (35)	224 (49)	228 (36)	232 (44)
10			247 (40)	246 (36)

*NWEA percentile score (based on 2002 norms study) is in parentheses

It is desirable for proficiency standards at the upper grades to be calibrated with standards in the prior grades. By that we mean that the proficiency standard at grade 4, should not be substantively more or less difficult to attain, in relative terms, than the proficiency standard for grades 8 and 10. In mathematics, the 4th grade cut score is set considerably higher than the cut scores for grade 8 and 10. If, in fact, fewer 4th graders achieve proficiency on the MontCAS than students at other grades, the higher cut score may create a misconception that the difference is due to deficiencies in 4th grade curriculum or instruction. In this case, our data would suggest that the higher cut score established as the grade 4 standard would be at least a part of the explanation for such a difference.

In reading, the problem is reversed, although the difference is not nearly as large. Based on our data, the 8th grade reading cut score is moderately more difficult, relative to our norm population, than the 4th grade standard. This creates the risk that some students who are identified as proficient on the 4th grade test may, despite showing normal growth, be identified as only nearly proficient on the 8th grade test.

Figure 5 depicts these issues by showing the difference between the established cut scores and a calibrated cut score that reflects holding the standard constant based on the grade with the most challenging standard.

Figure 5 -

Projected RIT cut score for MontCAS vs. Calibrated RIT cut score
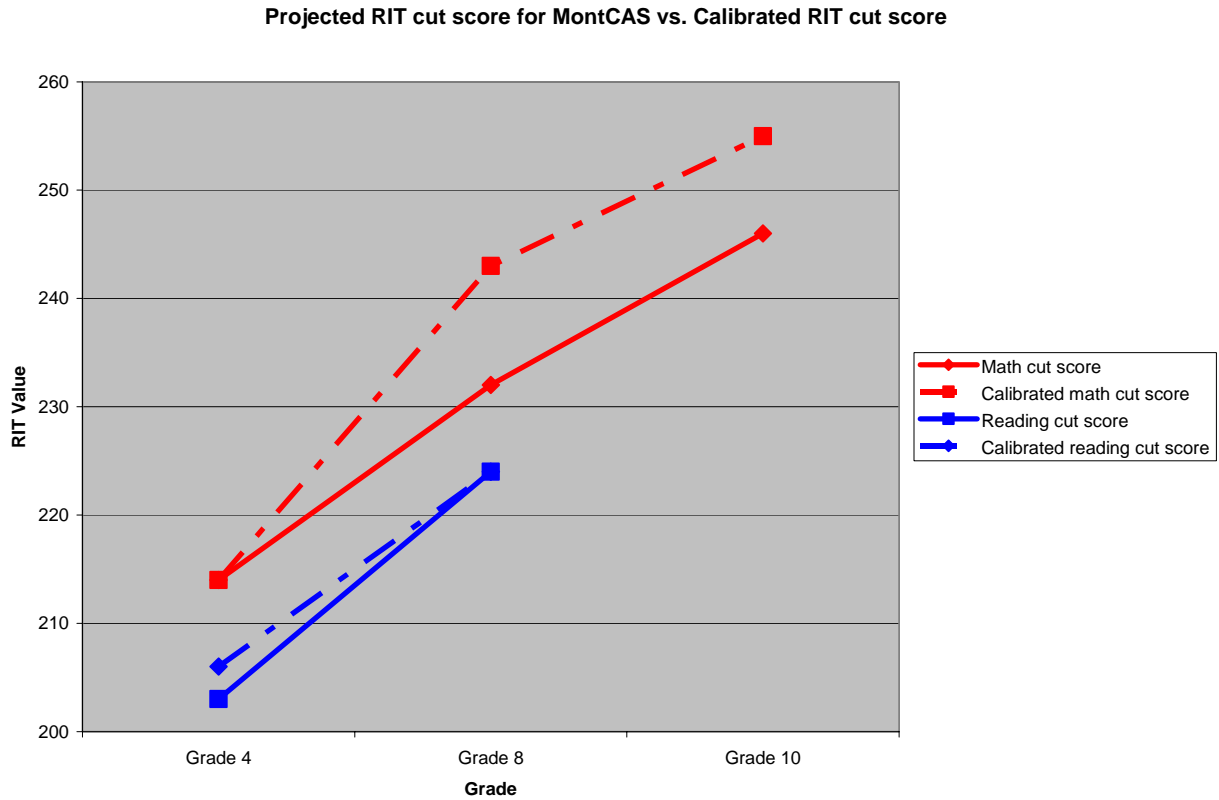
Comparing the MontCAS standards relative to those in place in other states

Northwest Evaluation Association tests have been aligned with the cut scores state assessments in 16 states. To get an estimate of the difficulty of the MontCAS 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.

The results are summarized in tables 11 and 12. With the 2004 adjustment in cut scores, Montana's standards now typically fall in the middle to upper-tier relative to the other states that we've studied.

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.

Table 13 - Cut scores representing "proficient" or "meets standards" level of performance on 16 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
NV	202	58	WY	214	73	SC	218	68	SC	222	64	SC	226	67	WY	232	74	IA	224	43	OR	236	77
CA	200	51	SC	209	59	NV	215	59	CA	216	46	CA	221	50	SC	230	68	ID	221	37	ID	224	44
SC	196	42	CA	205	46	CA	214	56	ID	211	35	WA	219	46	OR	227	58	CO	204	9	IA	223	44
MN	196	42	MT 04	203	42	PA	212	50	IN	210	32	IA	216	37	CA	226	54				WA	220	35
OR	193	35	ID	200	32	AZ	210	45	IA	209	30	NV	215	35	AZ	224	49				CO	209	15
ID	193	35	WA	199	32	OR	209	42	TX	208	28	ID	215	35	MT04	224	49				SC	209	15
IL	193	35	MT 03	196	26	IL	207	37	CO	197	11	TX	210	24	PA	223	46				CA	208	14
IN	192	32	IA	196	26	MN	207	37				CO	206	18	IN	219	35						
IA	191	31	NV	194	22	ID	206	35							IA	219	35						
AZ	190	29	CO	191	18	IA	205	32							MT03	219	35						
TX	179	13				TX	204	30							ID	218	32						
CO	179	13				CO	197	18							IL	218	32						
															MN	218	32						
															CO	206	12						

In South Carolina and California the standard reflects the performance level required as a prerequisite to graduation.   In Colorado, we report the standard used for NCLB proficiency

Table 14 - Cut scores representing "proficient" or "meets standards" level of performance on 16 state assessments - 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	212	84	WY	221	83	SC	230	81	SC	232	73	SC	241	76	WY	257	89	IA	241	44	MT 02	247	40
CA	204	63	SC	219	78	CA	225	71	CA	230	68	CA	238	71	AZ	248	75	ID	240	42	IA	247	40
NV	203	59	MT04	214	66	AZ	220	59	IN	221	47	WA	235	65	SC	247	73	CO	235	32	MT04	246	36
IN	201	50	CA	212	59	NV	216	48	ID	219	42	ID	225	44	CA	240	60				OR	245	33
MN	200	49	WA	210	54	PA	216	48	IA	218	40	IA	222	38	PA	237	53				ID	242	25
OR	199	46	ID	205	39	OR	215	46	CO	207	19	TX	221	35	OR	235	50				WA	242	25
AZ	199	46	IA	205	39	ID	213	41				NV	220	33	ID	233	46				CO	233	14
IA	197	39	MT 03	205	39	IA	212	38				CO	216	26	MT04	232	44				CA	232	13
ID	196	36	NV	200	26	MN	211	36							MN	231	42				SC	223	7
IL	193	29				IL	210	33							IN	231	42						
						TX	209	31							IL	230	40						
						CO	201	15							MT03	228	36						
															IA	228	36						
															CO	225	31						

Summary and Conclusions

This study investigated the relationship between the scales used for the MontCAS 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 MontCAS performance levels in those subjects. Test records for more than 4,000 students were included in this study.

Three methods generated an estimate of RIT cut scores that could be used to project MontCAS performance levels. Rasch SOS methods generally produced the most accurate cut score estimates. Accuracy of predicting MontCAS 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 Montana educators use NWEA assessments to better inform, plan and deliver student instruction. Good information, when matched with the professionalism and commitment of our Montana colleagues, will assure that every student has the opportunity to reach their aspirations.

References

Cronin, J., and Kingsbury, G. (2003). Alignment of the Iowa Test of Basic Skills with the NWEA RIT Scale – Montana. 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.