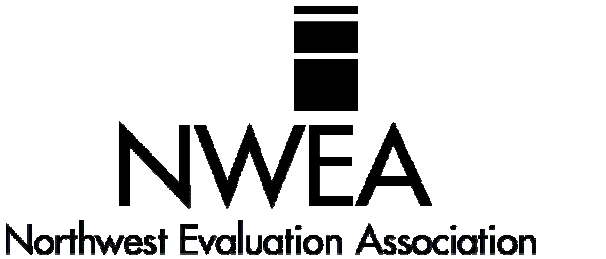
Assessing changes in the projected NWEA RIT scale cut scores for the 2002 and 2004 study of alignment with the Palmetto Achievement Challenge Tests

John Cronin, Ph.D and Martha McCall, Ph.D.

August 2004



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

Assessing changes in the projected NWEA RIT scale cut scores for the 2002 and 2004 study of alignment with the Palmetto Achievement Challenge Tests

John Cronin and Martha McCall

Septebmer, 2004

Northwest Evaluation Association regularly conducts studies of alignment between the RIT scale and the scales used for statewide achievement tests. We have conducted studies of alignment between the NWEA RIT Scale and 17 state assessments to date. In several of the states we serve, we are beginning to undertake follow-up studies to monitor the ongoing alignment between the RIT scale and state achievement examinations. Our recent follow-up study of the Palmetto Achievement Challenge Tests (PACT) is one of the first of these studies, using data from the spring 2002 and spring 2004 administration of these tests.

In recent years, educational experts have put increasing emphasis on the need for triangulation of student achievement information in order to assure that important educational decisions are based on data that is robust and corroborated. Indeed, many school systems use NWEA assessments as one way in which they triangulate data from their state assessment and other tests that may be in use in their district.

We strongly believe in the concept of triangulation. When a school learns that scores have not only declined on their state mathematics test, but also on other measurements that assess similar constructs, they have powerful evidence that the decline in state performance is not merely a statistical quirk. A sense of greater urgency emerges because the decline in performance reported by all of these assessments is likely to be real. There is also much that can be learned when triangulated assessments do not move in concert. Sometimes an improvement in a reading fluency assessment, for example, may not be reflected in the state measure of reading comprehension or in other measures in use in a school. When this is the case, teachers seek to learn the source of the difference to improve instruction. Often the very differences among a group of similar assessments provide the data that is key to improving instruction.

It is also possible that changes in the design or implementation of the tests themselves that may contribute to discrepant results. When this is the case, changes in student performance on one assessment may not reflect meaningful changes in student learning. Since accurate assessment of learning is NWEA's area of specialization, our organization has a special obligation to attend to issues in the design of tests that may lead to inconsistency in they way assessments perform relative to each other. That is why we try to carefully align our content with the state curriculum standards used by our members and why we undertake studies to determine the score points on our scale that align with the cutpoints for state assessments.

The final argument for triangulating assessment results comes from recognizing that even the best assessments do not measure learning with the same precision that an atomic clock measures time. And even the best students do not perform in perfect "flow" with their capacity on each day that an assessment is administered. So we also triangulate assessments in order to mitigate some of the effects stemming from the error of measure on a test or variance in the performance of a student during any single measurement event.

Unfortunately, some of the regulatory requirements increasingly associated with testing seem to demand a level of precision from tests that is not possible to attain using conventional technologies. Adequate Yearly Progress requirements are the most obvious example of this

problem. These requirements presume that we can assess performance with such precision that a school or school subgroup achieving only a 2% improvement in its proficient population may be sufficiently distinct from a similar school that achieved a 3% improvement to merit imposition of sanctions. The design of the Adequate Yearly Progress regulatory requirements tend to magnify the effect that even small changes in tests or testing conditions may have assessment performance. For example, a three day school closure due to a snowstorm would normally have no lasting affect on any students' learning. But such a closure, if it occurred a week before testing, might contribute to a slight depression in performance that would disastrously effect a few school's AYP results. Even a slight decline in scores might cause as many as 3 to 6% of the student population to fall below a proficiency bar. In the measurement arena, a change in scale conversions that resulted in student performance estimates being rounded down rather than up, could also cause as many as 3 to 6% of the student population to move below a proficiency bar. Both these events would detrimentally affect the status of the school, but neither reflects a meaningful change in student achievement, or a serious flaw in test design.

We make this point regarding AYP both to illustrate why schools and test publishers are rightfully concerned about the consistency of their various assessment results and also why the AYP rules may have been adopted with an assumption that a level of consistency and accuracy is possible that cannot be realistically achieved. Nevertheless, this does not make it less important monitor and report test consistency.

Differences in the 2002 and 2004 projected NWEA proficiency level estimates

The catalyst for this investigation was completion of a recent study to confirm and monitor the alignment between the PACT and NWEA assessments (Cronin, 2004). For the most part we found that the two assessments remained closely correlated and that NWEA results predicted PACT status reasonably well. Nevertheless, we also found that performance level estimates at a few grades showed large changes and that the predictive accuracy statistics stemming from the 2004 study were both lower than those generated from the 2002 study and also lower than those gathered from most of our other state studies.

Tables 1 and 2 show the differences in estimated cut scores.

Of particular concern were large differences in estimates of the proficient cut score estimate for grade 3 relative to the PACT English/Language Arts assessment and the estimates for grades 3, 7, and 8 relative to the PACT mathematics assessment. The 2004 study generated proficient level estimates that were 7 RITs lower than the corresponding 2002 estimates for grade 3 reading. The 2004 study also generated mathematics proficient level estimates that were 4 RITs higher for grade 3, and 4 points lower for grades 7 and 8. The differences in percentile scores associated with these changes show that they affect significant numbers of students. The 7 point decline in the 2004 estimate, for example, would project to an estimate of an additional 19% of the NWEA norm population achieving above the proficient bar.

Table 1 – Differences between projected Reading RITperformance level estimates relative to the PACT ELA for 2002 and 2004*

	Basic			Proficient			Advanced		
Grade	2002	2004	Diff	2002	2004	Diff	2002	2004	Diff
3	187 (23)	182 (16)	-5	203 (61)	196 (42)	-7	217 (93)	211 (82)	-6
4	197 (28)	194 (22)	-3	212 (67)	209 (59)	-3	225 (94)	226 (95)	+1
5	208 (40)	202 (26)	-6	220 (73)	218 (68)	-2	231 (94)	232 (95)	+1
6	209 (30)	210 (32)	+1	221 (63)	222 (66)	+1	232 (90)	233 (91)	+1
7	210 (23)	210 (23)	0	226 (67)	226 (67)	0	236 (90)	238 (93)	+2
8	217 (30)	213 (22)	-4	230 (68)	230 (68)	0	237 (86)	240 (91)	+3

* associated percentile score from 2002 NWEA norms study is in parentheses

Table 2 – Differences between projected Mathematics RIT performance level estimates relative to the PACT for 2002 and 2004*

	Basic			Proficient			Advanced		
Grade	2002	2004	Diff	2002	2004	Diff	2002	2004	Diff
3	192 (26)	193 (28)	+1	208 (75)	212 (84)	+4	217 (92)	220 (95)	+3
4	203 (33)	202 (31)	-1	217 (74)	219 (78)	+2	228 (92)	228 (92)	0
5	210 (33)	212 (38)	+2	227 (76)	229 (80)	+2	236 (90)	236 (90)	0
6	215 (34)	215 (34)	0	234 (77)	232 (73)	-2	246 (92)	240 (86)	-6
7	224 (42)	223 (39)	-1	242 (78)	238 (71)	-4	251 (90)	247 (85)	-4
8	228 (37)	228 (37)	0	251 (80)	247 (73)	-4	262 (94)	256 (88)	-6

* associated percentile score from 2002 NWEA norms study is in parentheses

Investigation of sources of difference

Consistency in methodology

NWEA uses a consistent, standardized methodology for conducting our alignment studies. This includes assembling a study population that is large enough to provide stable estimates with performance distribution adequately spread across all state test performance levels to provide robust estimates of cut scores. This also includes employing three statistical methods for estimating cut scores. This provides the best assurance that the model used for prediction is the one that most closely reflects the statistical relationships between the two assessments. We also publish estimates that show the accuracy with which our test predicted performance with this state assessment relative to the estimates achieved in other studies. Rather than repeat all the details about the study methodologies here, we refer readers to the companion document to this

paper, which is the most recently published PACT study, for a complete discussion (Cronin, 2004).

Although we do standardize our methodology, we do not attempt to use identical study populations for these kinds of follow-up studies, since we can generally get a larger and more complete representation of our member districts to participate in a follow-up study. The 2002 study included over 11,500 students enrolled in grades 3 through 8 in two South Carolina school systems. The 2004 study included over 22,000 students enrolled in grades 3 through 8 in three South Carolina school systems. One of the school systems participated in both the 2002 and 2004 studies. Differences in study populations are one factor that could contribute to differences in RIT score estimates of PACT performance levels, although both study populations shared all the performance characteristics needed to obtain good cut score estimates.

Removal of out-of-level test records from 2002 data set and the effect on inter-test correlations and estimated cut scores

For this investigation we also reinspected the data used for the 2002 study and discovered that a small number of student records were included that reflected out-of-level testing on the PACT. Because the intention of the study was to establish alignment between the NWEA scale and the various grade level versions of the PACT, it was obviously important to remove student scores which were not achieved by taking that form. We removed these records and recalculated our projections of cut score estimates for the proficient level of performance on the 2002 study. Table 3 shows the number of records removed by grade for out-of-level testing.

Table 3 – Records removed from 2002 data because PACT was delivered out-of-level

Grade	Records Removed ELA	Records Removed Mathematics
3	0	0
4	4	0
5	8	3
6	2	0
7	2	4
8	0	2

Tables 4 and 5 show the revised estimates for the proficient level of performance. For the PACT ELA test, the removal of records resulted in a 2 point increase in the 2002 proficiency score estimate for grades 5 and 7. No change occurred in any of the other grades.

Table 4 – Differences between projected Reading RIT cut scores relative to the PACT ELA for 2002 and 2004

	Proficient		
Grade	2002	2004	Diff
3	203 (61)	196 (42)	-7
4	212 (67)	209 (59)	-3
5	*222* (77)	218 (68)	-4
6	221 (63)	222 (66)	+1
7	*228* (73)	226 (67)	-2
8	230 (68)	230 (68)	0

Score changes from the 2002 study are in italics

Table 5 – Differences between projected Mathematics RIT cut scores relative to the PACT for 2002 and 2004

	Proficient		
Grade	2002	2004	Diff
3	208 (75)	212 (84)	+4
4	217 (74)	219 (78)	+2
5	*228* (78)	229 (80)	+1
6	234 (75)	232 (73)	-2
7	*241* (76)	238 (71)	-3
8	251 (80)	247 (73)	-4

For the PACT mathematics assessment, estimated proficiency levels changed by one point at grade 5 and grade 7, in both cases reducing the difference between the 2002 and 2004 estimates. The inclusion of these out-of-level records had served to slightly reduce the correlation coefficients between the 2002 PACT ELA and NWEA assessments. The removal of records did not change the correlation coefficient estimates for mathematics. Table 6 reports the revised same subject correlation coefficients and notes changes the size of changes.

Table 6 – Inter-test correlations for PACT and NWEA assessments by subject

	NWEA Reading to PACT ELA		NWEA mathematics to PACT mathematics	
	2002	2004	2002	2004
Grade 3	.77	.76	.77	.76
Grade 4	.78 (+.02)	.79	.85	.84
Grade 5	.76 (+.06)	.78	.84	.84
Grade 6	.77	.78	.87	.84
Grade 7	.79 (+.01)	.78	.85	.85
Grade 8	.81	.76	.85	.85

change in 2002 coefficients due to removal of out of level tests are noted in parentheses

Questions related to content alignment

The NWEA assessments used in South Carolina are also designed to align closely with the state's content standards. NWEA assessments measure reading and language usage as separate disciplines, however, while the PACT combines the assessment of reading and writing in a single score reflecting performance in English/Language Arts (ELA). This assessment also includes a performance writing component that is not offered on the NWEA language usage assessment. Thus, while the NWEA reading and language usage assessments do provide a closely aligned assessment of the content standards in these subjects, neither the NWEA reading assessment nor the language usage assessment alone covers the complete ELA domain. It is possible therefore, that changes in cut score estimates might occur if student performance on the portion of the ELA not covered by an NWEA test also changed.

NWEA's 2002 study included only our reading and mathematics tests because the participating school systems had not implemented tests of language usage at that time. The 2004 study included reading, language usage and mathematics. Rather than attempt to combine NWEA reading and language usage scores in some fashion to generate a combined estimate of ELA performance, we chose to align the reading and language usage separately. This results in slightly lower correlation estimates than are usually achieved when the state assesses reading and writing separately. We do recommend that school systems use the results of the reading and language usage tests in concert when making instructional decisions in the literacy domain.

Differences in the performance characteristics of the 2002 and 2004 study populations

In general the 2002 and 2004 study populations exhibited similar performance characteristics relative to each test. Tables 3 and 4 show the relevant univariate statistics reflecting reading and mathematics performance of the study populations on the 2002 and 2004 assessments. In English/Language Arts the means on PACT and NWEA assessments of the two study groups for the two years were similar except for grade 3. In grade 3, the 2004 PACT mean was 7.04 scale points higher than the mean for the 2002 study group. On the NWEA assessment, however, the mean score for reading was 1.41 points lower than the 2002 study group. The NWEA reading assessments for both the 2002 and 2004 groups show higher negative skew than the results for the PACT ELA. There are a number of reasons why this might occur. We believe the most likely reason for the difference in skew is that the NWEA assessments are designed to adapt item difficulties to the learner and are not constrained by the requirement that they offer items that are focused near the level of the performance standard for a particular grade. This typically gives the

NWEA assessments greater range when measuring low performing students. The other noticeable difference was a change in the skew statistic for grade 6. The distribution of the 2004 study group on 6th grade PACT ELA was slightly negative (-.155). This was a difference of (+.266) from the 2002 skew statistic (-.421). The skew for the two 6th grade samples on the NWEA test moved slightly in the opposite direction (-.442 to -.544), with 2004 group showing a slightly larger negative skew (-.102).

Table 7 – Mean, Median, and Skewness of PACT ELA and NWEA reading assessments

	2002 PACT ELA			2004 PACT ELA			2002 NWEA Reading			2004 NWEA Reading		
Grade	Mean	SD	Skew	Mean	SD	Skew	Mean	SD	Skew	Mean	SD	Skew
3	**309.09**	**13.17**	**-.109**	**316.13**	**14.90**	**-.196**	**201.40**	**12.28**	**-.625**	**199.99**	**12.37**	**-.613**
4	407.55	11.80	-.084	408.33	12.72	-.171	209.31	10.90	-.501	207.16	11.43	-.551
5	**504.31**	**12.22**	**.022**	**506.69**	**13.12**	**-.245**	**214.29**	**10.44**	**-.477**	**213.00**	**11.17**	**-.546**
6	609.32	16.00	-.421	605.95	16.07	-.155	218.86	11.10	-.442	216.61	11.69	-.544
7	706.88	13.51	-.099	706.62	13.33	-.094	221.48	11.22	-.527	220.19	11.62	-.538
8	806.35	13.79	-.328	807.42	12.26	-.110	224.58	10.70	-.515	223.52	11.40	-.625

* Bolded text indicates grades in which adjusted cut point estimates of 4 points or greater were found

Table 8 – Mean, Median, and Skewness of PACT ELA and NWEA mathematics assessments

	2002 PACT Mathematics			2004 PACT mathematics			2002 NWEA Mathematics			2004 NWEA Mathematics		
Grade	Mean	SD	Skew	Mean	SD	Skew	Mean	SD	Skew	Mean	SD	Skew
3	**310.18**	**14.91**	**-.095**	**313.02**	**12.43**	**-.206**	**202.69**	**10.87**	**-.100**	**207.07**	**11.69**	**-.095**
4	413.50	15.52	-.173	414.70	14.66	-.054	214.32	12.13	.117	216.99	12.62	-.194
5	510.91	16.35	-.126	514.80	15.47	-.175	222.39	13.42	-.179	225.72	12.89	-.365
6	611.26	15.82	-.150	615.58	14.95	-.423	228.30	15.20	.002	229.79	13.59	-.264
7	709.25	16.22	.039	713.66	15.40	-.119	232.56	15.36	-.216	234.83	14.09	-.447
8	**807.23**	**13.94**	**.004**	**810.53**	**13.88**	**-.099**	**237.32**	**15.81**	**-.272**	**239.56**	**14.53**	**-.307**

* Bolded text indicates grades in which adjusted cut point estimates of 4 points or greater were found

The differences in the two grade 3 reading cut score estimates can now be partially explained by a difference in performance that occurred between the groups on one measure, the PACT ELA, that was not reflected in the other measure, the NWEA Reading assessment. Such a difference would cause a reduction in the estimated cut score. While this explains the statistical phenomenon, it really doesn't help explain the underlying cause.

Interestingly, the other grade levels that experienced large differences do not show similar discrepancies in mean scores that would serve as obvious explanations for the change. Nor were there clearly visible changes in the skew statistics that would help offer explanation.

We did find that the 2004 PACT and NWEA mathematics scores were consistently higher, by about 2 to 4 points, than the scores for the 2002 study population. Because the direction of these differences was the same on each test, it is not likely that the higher performance achieved by the 2004 population would be a factor that would have great influence on changes in the projected proficiency scores.

How small changes in performance, alignment, or scales may cause large changes in distributions of students

With the exceptions that we have already noted, we did not find large differences in the performance of the 2002 and 2004 study populations on either the PACT or NWEA measures. We did discover, however, that large portions of the study populations clustered near the

proficiency level cut scores on both the PACT and NWEA assessments. Tables 9 through 12 summarize these distributions.

Table 9 – Proportion of study population on or near PACT proficiency cut scores – 2004 ELA

	Cut Point	% on cutpoint	% on or within 1 scorepoint	% on or within 1 rounded SEM
Grade 3	310	2.3	7.8	28.4
Grade 4	410	4.8	9.9	38.1
Grade 5	511	2.9	10.3	34.5
Grade 6	612	3.0	9.9*	22.3
Grade 7	712	2.8	9.4	25.5
Grade 8	813	2.5	9.0	23.4

* The cutpoint for grade 6 was 612, the nearest adjacent score on the PACT scale was 610 in 2004

Table 10 – Proportion of study population on or near RIT estimated proficiency cut scores – 2004 NWEA Reading

	Estimated Cut Point	% on cutpoint	% on or within 1 scorepoint	% on or within 1 rounded SEM
Grade 3	196	2.9	8.3	17.6
Grade 4	209	3.4	11.0	22.1
Grade 5	218	3.3	10.4	21.1
Grade 6	222	3.3	9.3	14.8
Grade 7	226	3.6	9.9	19.9
Grade 8	230	3.5	10.6	19.9

Table 11 – Proportion of study population on or near PACT proficiency cut scores – 2004 Mathematics

	Cut Point	% on cutpoint	% on or within 1 scorepoint	% on or within 1 rounded SEM
Grade 3	317 *	5.5	17.6	33.6
Grade 4	416*	4.1	12.5	28.1
Grade 5	517	3.7	11.5	20.0
Grade 6	618*	3.7	10.0	23.4
Grade 7	717	2.9	9.9	24.6
Grade 8	818	2.3	6.6	21.8

* Reflects the first actual score awarded proficiency status because scores on cut point was not awarded.

Table 12 – Proportion of study population on or near estimated RIT proficiency cut scores – 2004 NWEA Mathematics

	Cut Point	% on cutpoint	% on or within 1 scorepoint	% on or within 1 rounded SEM
Grade 3	212	3.3	9.4	18.1
Grade 4	219	2.8	8.7	16.8
Grade 5	229	2.8	8.5	17.4
Grade 6	232	2.2	8.1	17.0
Grade 7	238	3.1	9.9	17.6
Grade 8	251	2.4	7.3	13.0

Distributions of this kind mean that proficiency level statistics can be dramatically influenced by small events. A simple change in the way a scale score rounds may move 2 to 5% of students above or below a cutpoint. Changes in test administration dates, even if they have only a one point affect on performance, would cause substantive changes in proficiency level performance. Because small changes in scores make big changes in proficiency level distributions, it is obviously in the interest of states to monitor conditions related to testing and methodologies used for scaling with vigilance.

Several of the differences found between the 2002 and 2004 studies were large. Our organization is not in a position to be able to confirm the reasons for the cut score estimate changes in PACT relative to our test. The most obvious reason is that the 2002 and 2004 tests studied involved students who cannot be expected to perform with perfect consistency every time. Because those being measured are not entirely consistent, it is impossible for two scales to perfectly calibrate their performance by measuring targets that, while generally consistent, nevertheless do move. In the case of the NWEA reading and PACT ELA assessments, we also know that the tests measure closely correlated but not identical content. One possible reason for the differences in the Grade 3 estimates for these subjects could be related to changes in performance or design of the language and writing portion of the ELA assessment that would not be picked up by a reading test. Indeed, NWEA administers a language usage assessment separately because the constructs that predict writing performance have some fundamental difference from reading.

We are also aware that the 2002 and 2004 versions of PACT ELA had differences in content which may affect the way the two scales track. In particular, the removal of writing tasks that were based on reading prompts, which seems to us a prudent design choice, may have had an effect. The choice was prudent because some students may have also been penalized for being unable to demonstrate a writing skill when they could not perform entirely because they didn't correctly understand the reading portion of the question. If some student's scores were unfairly deflated because of this design issue, it is entirely possible that, when corrected, student performance on the two tests would calibrate differently.

Finally, there is also the possibility of unexpected changes in the test scale that may inadvertently influence the proficiency level. All test publishers, including those responsible for PACT, take elaborate measures when constructing tests to minimize the possibility of drift in the underlying scale. For PACT, in the past these strategies included careful field testing of new items and anchoring the difficulty of new test items to questions that have appeared and performed consistently on the previous version of the test. No methodology produces perfectly calibrated measurement, however. Even the most precise of clocks do not measure time with absolute accuracy. Because so many student scores are concentrated within one standard error of measure of the performance bar, seemingly minor inconsistencies in the performance of measures may have large effects. We mention this not to criticize the methods of scaling employed on the PACT, but to encourage ongoing review, monitoring, and adjustment of these processes as necessary. Our hope is that providing external information to triangulate state testing data, we can be of some service to our colleagues in state departments of education as they do this work. Testing remains a very human science, and even when the best methodologies are conscientiously employed to most tests in use, calibrating scales so that they are rock-solid identical in their true difficulty from year to year may be asking more of scaling technologies than is currently possible.