Alternate Methodologies for Estimating State Standards on a Widely-Used Computer Adaptive Test

John Cronin, G.Gage Kingsbury, Michael Dahlin, Deborah Adkins,

Northwest Evaluation Association

Branin Bowe, National Heritage Academies.

Alternate Methodologies for Estimating State Standards on a Widely-Used Computer Adaptive Test

John Cronin, G.Gage Kingsbury, Michael Dahlin, and Branin Bowe

The No Child Left behind Act implements a federally mandated accountability system in which state assessments are the cornerstone measure of student achievement and progress. For the vast majority of schools, the state assessment is their single most important measure of student achievement and for the public, results of state assessments may be treated as the primary evidence of the effectiveness of their schools. Because the Act explicitly requires that 100% of students demonstrate proficiency in reading and mathematics by 2014, all states use the percent of students reported proficient or better as their primary metric related to educational achievement.

Because each state sets its own standard, it is impossible to compare student achievement across states by directly comparing the percent of students reported as proficient on these tests. Nevertheless, there are compelling reasons to attempt comparisons among tests. In particular, as the time approaches for the reauthorization of the Act, the public has an interest in knowing the relative rigor of the standards states have implemented and how to interpret student performance in light of the rigor of each standard. One possible way to accomplish this is by attempting to link state assessments to a common scale that could be used to evaluate the rigor of standards and provide a common ruler for measuring student performance.

Debates about test and scale linking have raged since the middle 1960's within the psychometric establishment. Linn (2005) notes that the lead articles in the very first issue of the Journal of Educational Measurement are all related to equating issues (Angoff, 1964; Flanagan, 1964; Lennon, 1964; and Lindquist, 1964). As more states began to implement their own systems of standards and assessments, interest in equating these examinations to other assessments, particularly NAEP, emerged. This effort first dates from Kentucky's project to report Kentucky Instructional Results Information System test results to the NAEP scale (Kentucky Department of Education, 1993). This was followed by several other efforts to link NAEP to other state assessments and nationally published tests throughout the 1990's that are ably reviewed by Thissen (2005). These results of these initial efforts were greeted with skepticism. Based on their own investigation and evidence gained from this body of prior work, a National Research Council Committee chaired by Paul Holland concluded that the assessments of different states varied too much in content, item format, conditions of administration, and stakes to justify equating the tests on a single scale (Feuer, Holland, Green, Bertanthal, & Hemphill, 1999). Their conclusion was that state tests

could not be *equated* to NAEP, a finding which defused efforts to replace NAEP by linking the various state assessments to a single common scale.

Since the introduction of the No Child Left Behind Act, however, researcher interest in investigating the rigor of the state proficiency standards that define quality learning in the act has sustained interest in this topic, not for purposes of using a common scale to replace a test, but instead in order to find reasonable means to compare state standards and student performance relative to them. In this light, Dorans and Holland's (2000) standards for *equating* seem overly strict because they demand that the tests being linked measure equal constructs, with equal reliability, in a symmetrical fashion, with no population variance among the groups measured. Tests that meet these criteria would be duplicative and one could legitimately argue that rather than link the two tests, it would be better if we simply chose one and gave it to everybody (which was the proposal that the NRC rightly rejected).

But we rarely give two math tests to the same group of students for the purpose of giving them redundant work. Typically we give multiple tests of the same domain because they serve unique purposes, and these purposes will require differences in design that inevitably violate the equating criteria. For example, there are excellent reasons why one might want to administer a state mathematics proficiency test and a test measuring college readiness in mathematics to many of the same students. These tests are likely to measure different constructs, have differing reliability coefficients and not produce symmetrical results. Nevertheless, there remain good reasons that to perform *predictive* linkages between these assessments. Such linkages might address such questions as: Is proficiency on this state test equivalent to the performance needed to demonstrate college readiness? Have improvements in proficiency been matched by improvements in college readiness? These are important questions and linking strategies can be used to inform them.

Prediction, as described by Mislevy (1992) and Linn (1993) does not necessarily require that the linked assessments be so similar the constructs measured, in design, and in the distribution of results that the tests could *equate* to or replace one another. Instead, *prediction* demands that the test measure the same general domain in a highly correlated fashion, and that the aggregated results of a linked assessment project to highly similar aggregated results for the companion assessment. Differences in item format, construct, or test design are acceptable if they do not unduly degrade the projection. This is the basis Linn

uses to call for more work to provide a quantitative basis for evaluating the quality of linkages and such research would be very desirable.

Pursuant to this logic, McGlaughlin and Bandiera de Mello (2002, 2003) constructed predictive linkages between a number of state assessments and the 2000 and 2002 administrations of NAEP using an equipercentile methodology with school level data. Their methodology produced reasonable standard errors of measures for the most part, although the error of measurement increased when attempting to link the lowest level of performance on some of the state assessments, a problem which may be a function of differences in the measurement ranges of NAEP and the companion state assessments. Their estimate of state proficiency standards found very wide variations in proficiency standards among the 29 states they studied.

Braun and Qian (2005) refined McGlaughlin and Bandiera de Mello's methodology by including NAEP's sampling weights and jackknife procedure in their estimates to produce more robust aggregated results. This approach called "weighted aggregate mapping" produced results similar to the McGlaughlin and Bandiera de Mello's method when used side-by-side. Braun and Qian's results also confirmed large differences in the estimates of state proficiency standards.

Kingsbury, Olson, Cronin, Hauser, and Houser (2003) produced estimates of proficiency standards for fourteen states using pools of students who had taken both their state proficiency examination and Northwest Evaluation Association's Measures of Academic Progress (MAP). They applied linear regression, quadratic regression, and a Rasch Status on Standard (SOS) methodology to generate three estimates of cut scores, using the estimate that produced the most accurate level of prediction and lowest rate of Type I error to render an official estimate of difficulty. Their approach addressed four common difficulties in estimates derived from NAEP. First the estimates used students who were known to have taken both assessments, thus avoiding the issues inherent in trying to create equivalent population samples, when one of the tests does not report results at the student level. Second the NWEA measure was aligned to the content of the state performance standards, thus reducing the error due to differences in measurement construct in relation to NAEP. Finally, because the measure is adaptive, the assessment's design targets the bulk of questions to the student's level of performance. This reduces error that may be a product of the need to construct fixed form tests so that most items are focused at or near the level designated as proficient. Fourth, because the NWEA tests are offered at grades 3 through 10, their

analysis included permitted estimates in states that did not offer state testing in the NAEP grades. It also allowed the researchers to investigate issues related to the calibration of standards across a range of grades.

Kingsbury's group replicated prior findings about the variance in state proficiency standards. They also raised questions around the calibration of standards across grades, finding that differences in performance between grades 3 and 4 and grades 7 and 8 within some states were more likely to be a product of differences in the rigor of cut scores than real differences in performance among these grades.

Collectively, these studies have provided credible evidence that standards of proficiency indeed vary a great deal across states, and that the forms of linking employed in these studies are methodologically adequate for that purpose. In other words, the methods employed to date are robust enough to demonstrate that standards vary and are adequate to document that the proficiency cut scores in the most stringent states are indeed more challenging than those in the least stringent states.

Thissen (2005), however, raises concerns around whether current methodologies systematically understate the rigor of state standards and whether the current approaches to linkage have the capacity to make fine distinctions in their difficulty. In particular, Thissen criticizes the precision of standards estimates derived from NAEP because of the differences in motivation that may exist between high stakes state assessments and the low-stakes NAEP assessment. If the difference in motivation were to affect the estimate, it would understate the academic performance of the NAEP test-takers and, as a consequence, artificially inflate the difficulty of the NAEP standard. If true, this may make state standards seem easier to achieve than they really are. While this does not materially affect our ability to compare state standards to one another, it might lead researchers to conclude that state standards are less rigorous than they may be in reality. Complicating the issue is the fact that state assessments themselves have wide variance in the stakes imposed on students. Some have significant consequences (or rewards) for students based on their personal performance, others have high stakes for schools and teachers but virtually no stakes for the test-takers themselves. Thus one cannot assume that the level of motivation associated with students on state tests is a constant across these 50+ venues.

Thissen also raised questions related to the stability of estimates over time, noting that these questions can be resolved empirically by producing multiple linkages among tests. Presumably this done to meet the expectation that linked scales meet the equating requirement of "invariance of results". This is an

important concern when one's objective is to produce linkages that will be used to project future results on one assessment to another instrument. However, many researchers are just as interested in knowing when relationships between linked assessments are **not stable**. Unstable relationships between two measures may indicate that proficiency standards have drifted, or indicate that improvements measured on one assessment are not sustained on other measures of the domain. The latter is particularly important because the underlying intention of the No Child Left Behind Act was not to produce improvements in state test results. Instead the Act was intended to ensure that all students, particularly traditionally underserved minorities, achieved a level of *proficiency* in reading, mathematics, and science that would ultimately improve many life outcomes. This is unlikely to occur if improvements on state assessments don't translate to other assessments in the same domain.

For example, the possibility that repeated use of a test form or repeated introduction of similar questions can render the initial calibrations of many items to a scale meaningless, because the distribution of student performance on these items may change due to increased familiarity and teaching to the test. This particular problem was first revealed by Linn, Graue, and Sanders (1990) who found that scores increase as a form is reused, particularly during the first few years. Over time, this can cause the pass point on one assessment scale to drift downward relative to the linked assessment. This kind of problem is often revealed when the improvement in performance achieved on the first test is not reflected in the second examination. While it is true in such instances that the predictive relationship between two such tests is not stable, identifying that fact is extremely useful to anyone who is interested in investigating whether improvements in student performance are the likely result of improved learning or instead from factors that may be unique to a particular exam. Koretz (2005) presents this same problem, finding that the idiosyncratic nature of item formats on state assessments may reduce the likelihood that these results generalize to other tests. Thus, evidence that scales may drift relative to one another is actually very useful and Thissen's proposed solution, producing multiple linkages among tests, would provide an excellent means of collecting this kind of evidence.

These problems are best document when one of the measures used in linking has demonstrated stability relative to the companion measure. The NAEP instrument used in the McGlaughlin and Bandiera De Mello and the Quin and Braun studies employs sophisticated procedures to ensure stability of the scale over time. The MAP instrument used by Kingsbury and colleagues also has demonstrated stability over the thirty year period of its use (Kingsbury, 2005).

Thus, despite the issues raised with past efforts to link scales among tests, there remain good reasons for attempt the work. Much of the past effort has used the NAEP scale to make linkages to the various state assessments. This is useful, but it is desirable to triangulate this work with research that develops linkages between other assessments and state exams. In particular, linkage using an assessment that is not as vulnerable to issues related to student motivation as NAEP would be useful.

In addition, because NAEP linkages must be nature be based on comparison of estimated school level results, approaches that use student level results for linkage or employ alternate methods for linking group results provide additional means to triangulate the findings coming from prior studies. Finally, because NAEP content cannot be individually aligned to the individual standards of the state, approaches that use assessments that are more closely aligned to state content standards might reduce some of the error associated with these estimates.

In that light, a study was undertaken to compare four methods to link state results to a computer-adaptive test, aligned to state content standards, that employs a vertical cross-grade scale for reporting results. Three of the methods used student level results to perform the linkage, while the fourth creates estimates from closely matched student populations. The test used was one that is used by school systems as part of their instructional program, with results generally reported to students and families. In addition, growth estimates from the assessment are often used as an accountability metric within these school systems.

Methodology

The study provided estimates linking proficient cut scores on state assessments in Arizona, Colorado, Delaware, Michigan, and New Hampshire to equivalent scores on a vertically scaled computer-adaptive assessment developed by Northwest Evaluation Association.

Instruments

Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP), a series of state aligned computer-adaptive assessments, were used to predict performance on the Arizona Instrument to Measure Standards (AIMS), Colorado Student Assessment Program (CSAP) tests, Delaware Student Testing Program (DSTP) assessments, New Hampshire's implementation of the New England Common Assessment Program (NECAP) and Michigan Educational Assessment Program (MEAP) tests. In general, linking was performed between the reading and mathematics score on the MAP assessments and

the state assessments' reported reading and mathematics score. For Arizona, however, the state reports an English/language arts score to meet the NCLB requirement, so this score was used for linking.

NWEA's MAP instruments are customized to align with each state's content standards. Generally this is accomplished by having NWEA curriculum experts perform a cross-walk between a detailed index used for items in the MAP item pool and the outline for the content standards given by the state. The results of this analysis are to select the items that will be included in the pool for the MAP assessment. From this pool a 40 to 55 item assessment is delivered to each student, with each standard assessed by an equivalent number of items to provide balance. This approach generally assures good content alignment between MAP and each state test, thus reducing the amount of error in a linkage estimate that might be attributed to differences in content. Curriculum alignment was not as close with the Arizona English/language arts assessment because the Arizona test introduced writing and usage components that are not measured by MAP reading.

MAP tests use a vertical, cross-grade scale called the RIT scale to measure and report student performance and growth over time. The original procedures used to derive the scale are described by Inglebo (1995). These procedures differ considerably from traditional methods of vertical scale construction that involve equating of different test forms. Some of the differences are as follows:

1. The entire MAP item pool is calibrated according to the RIT scale. There is no need to equate forms of tests, because each assessment is simply a subset of a single pool.
2. The original field test design for the paper version of MAP assured that each item was calibrated against items on at least 8 other forms of the paper test. This resulted in a very robust item pool with item calibrations that have remained largely constant for over 20 years (Kingsbury, 2003). The current field testing designs for the computer-adaptive version of MAP introduce two field test items into most student's MAP assessments. A minimum of 300 students take the item during the calibration process and because each student takes a uniquely assessment, each field test item is calibrated against a pool of 12,000 to 15,000 unique item responses.
3. Because the test is adaptive, students from a broad range of grades (ranging from 3 to 12) are eligible to view each item, both in field testing and in live tests. As a result, the vertical scale for the test is not extrapolated from equating versions of test forms intended for use within a grade across grades. Instead the scale is constructed from the ground up.

Because MAP tests are adaptive, low and high performing students will answer the questions that are most appropriate to their current level of achievement, but these will not always directly align to the standards for their particular grade. For example, a high performing 5th grade student in mathematics, may answer questions that are aligned with the states content standards at the 7th or 8th grade level.

All questions in the MAP tests offered in these states were multiple-choice in format.

In general, MAP tests are used by school systems for several purposes:

1. To provide immediate feedback (students receive results as soon as they complete the test) in regard to academic performance and provide information to teachers that guides instruction.
2. To inform placement decisions in advanced course work or gifted-talented programs. For struggling students, MAP results commonly inform placements into interventions or special programs.
3. To predict likely student performance on the state assessments.
4. As assessments to monitor student performance and growth for purposes of internal answerability or accountability.
5. As assessments to report individual student growth to parents and students.

While it is fair to say that MAP assessments are not typically taken under the same stakes as state assessments, MAP generally serves important purposes within the school systems that use it. As a result there are considerably stronger incentives to perform well on MAP than may exist for performance on NAEP assessments. While these incentives don't entirely address the issues around linking scales when students taking two tests experience different levels of motivation raised by Thissen (2005) and by prior work done by our organization (Cronin, 2005), they reduce the degree by which the two scales are effected by this factor.

Sample

Two samples of students were created for this study. The first sample (Group 1) was composed of students who took both their state assessment and the appropriate version of MAP. To accomplish this all NWEA member school systems within the state were invited to participate in a study to estimate state proficiency level cut scores on MAP. Those that volunteered to participate provided individual student results on the state assessment to our organization. Results of students with valid state test results were

linked to to their respective MAP data using name and Student ID. These matched records formed the sample for Group 1.

The second sample (Group 2) was created by finding a group of schools within each state in which similar numbers of students had participated in both MAP and state testing and using results from those schools. To form the group:

- All valid student test records for Northwest Evaluation Association clients in each state studied for the appropriate term were extracted and aggregated these results by school.
- Data was captured from each of the five states showing the number of students tested in each school and the proportion of students tested who performed at each proficiency level.
- NCES school identification information was used to link results from the state test to the appropriate school in the NWEA database.
- The dataset was filtered to find schools whose tested population on the NWEA assessment was between 95% and 105% of the population taking the respective state exam. If this method generated at least 1,000 students per grade, we did not expand the study group further. If the initial criteria failed to generate that number we liberalized the criteria to 92.5% to 107.5%. The 95% to 105% criteria worked in all cases with the exception of Arizona mathematics, Arizona reading, and Delaware reading. For these the more liberal criteria were applied.

Group 1 and group 2 students in Arizona, Colorado, and Delaware took NWEA tests between March 1st and June 15th of 2006 and their respective state examinations in March and April of 2006. Students in both study groups in New Hampshire and Michigan took NWEA tests between August 15th and December 1st of 2005 and their state assessment in October, 2005.

In general, the method used to create group 2 produced samples that reflected a larger number of school systems that we have been able to obtain through our other methodologies. The method did not, however, consistently produce larger samples of student records, primarily because of the attrition generated by the matching requirement.

Table I – Participants in the five linking studies

State	Regression/Rasch SOS (Group 1)				Distributional (Group 2)			
	Reading		Math		Reading		Math	
	Districts	Students	Districts	Students	Districts	Students	Districts	Students
Arizona	3	15,054	3	15,409	6 (32)	9,286	7 (36)	10,119
Colorado	6	20,924	6	19,885	38 (157)	32,192	40 (168)	36,823
Delaware	4	11,198	5	12,375	7 (13)	4,693	7 (35)	9,780
Michigan	5	14,942	5	14,875	23 (69)	16,775	17 (77)	12,275
New Hampshire	2	4,999	2	5,008	38 (74)	20,242	38 (75)	21,232

Generating Cut Score Estimates

Four methods were employed to create estimates of the score on the RIT scale that would be equivalent to cut score for proficient performance on each respective state test. Three of the methods employed were used in our prior study of state standards (Kingsbury et al, 2003). The most straightforward was Linear regression ($state_{pred} = a(RIT)+c$). Because departures from a linear relationship are often observed on the lower and upper end of state test scores, second-order regression ($state_{pred} = a(RIT^2) + b(RIT) + c$) was also employed. For each of these two methods, the RIT score equivalent to the state proficient score was estimated by substituting the appropriate state scale score for $state_{pred}$ and solving for RIT. A fixed-perameter Rasch model (Inglebo, 1997) was the third method employed. To derive this estimate, the proficient 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 tests. Student abilities, operationally reflected in the RIT score, were the fixed perameter used to anchor the difficulty estimate of the state-defined proficiency standard to the RIT scale. The resulting difficulty estimate was taken as the RIT cut score. This is referred to as the Rasch Status on Standard or Rasch SOS method. In our experience, it is sometimes more accurate than conventional regression methods when estimating cut scores near the low and high performance boundaries of a scale.

Linear regression, second-order regression and the Rasch SOS methodology were applied to group 1 data to create the proficiency score estimates for those methods. The fourth method, called a distributional method, generated an estimate of the proficient cut score by finding the proportion of students in group 2 who achieved proficiency on their state test, using an equipercentile method to estimate the score on the NWEA test that would generate the equivalent proportion of students.

We employed a Distributional method to generate proficient cut score estimates for group 1. We generated an estimate of the proportion of proficient students in group 1 by aggregating the state assessment results of the schools included in the sample. Then using student level NWEA results for the same schools, we found the performance point on the RIT scale that generated the same proportion of students.

Results

Pearson Correlations

To evaluate concurrent validity of the assessments, correlation coefficients for the student samples were generated between each of the five state assessments in Reading/English language arts and the equivalent version of MAP. Four of the five states test reading and these were correlated with the reading version of MAP. One state, Arizona, broadens their assessment to include language arts. We also correlated this test with MAP, although the state assessment is clearly broader in scope. Tables 2 and 3 show that average coefficient within the reading domain fell in a range of .76 to .82 by state. Correlations within the mathematics domain were stronger, ranging between an average of .82 and .86. Relationships between the mathematics tests were strong enough to suggest excellent predictive validity between MAP and the various state assessments. Relationships between the Arizona reading/ELA test and the Colorado test were similarly strong, but the strength of correlation between MAP and the reading assessments for Delaware, Michigan, and New Hampshire were less than ideal. Figures 1 through 3 provide examples of issues that may have contributed to the slightly lower correlations. All three figures show deterioration in the correlation near the bottom of the two test scales. In addition, the New Hampshire and Michigan data at these grades also show some evidence of floor effect, that is they show MAP with greater range of measurement at the lower end of the scale than seems to be present with the state assessment for that grade. This would not be surpising, since a test using an adaptive design should offer more accurate measurement at the ends of the performance continuum. For purposes of this study, these issues complicate efforts to predict cut scores when they are at the low end of the distribution, but should not influence the accuracy of cut scores that would fall in the middle of the distribution.

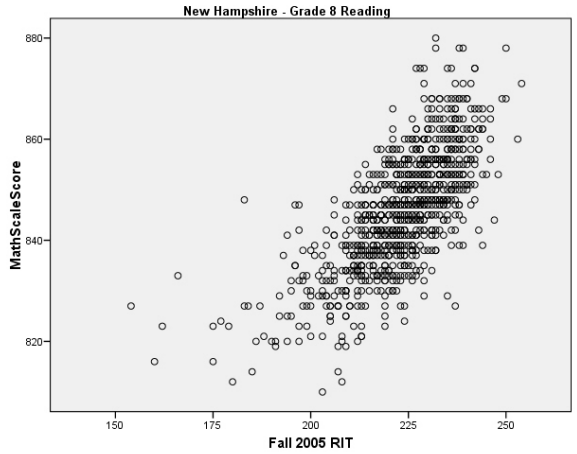
Table 2– Correlation between state reading/English Language Arts assessments and the equivalent version of MAP

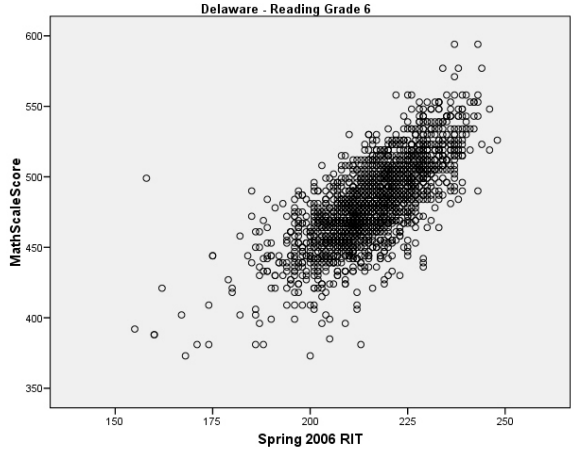
	Arizona ELA	Colorado	Delaware	Michigan	New Hampshire
Grade 3	0.85	0.79	0.76	0.76	0.82
Grade 4	0.82	0.83	0.76	0.78	0.79
Grade 5	0.83	0.83	0.75	0.77	0.74
Grade 6	0.82	0.83	0.74	0.77	0.79
Grade 7	0.81	0.81	0.78	0.75	0.79
Grade 8	0.80	0.81	0.78	0.77	0.71
Ave	**0.82**	**0.82**	**0.76**	**0.76**	**0.77**
Min	**0.80**	**0.79**	**0.74**	**0.75**	**0.71**
Max	**0.85**	**0.83**	**0.78**	**0.78**	**0.82**

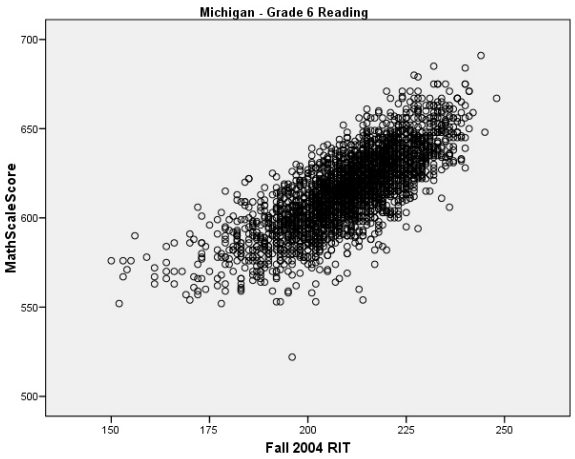
Table 3– Correlation between mathematics assessments and the equivalent version of MAP

	Arizona ELA	Colorado	Delaware	Michigan	New Hampshire
Grade 3	0.84	0.81	0.81	0.78	0.82
Grade 4	0.85	0.84	0.85	0.81	0.84
Grade 5	0.86	0.86	0.81	0.84	0.85
Grade 6	0.87	0.88	0.85	0.83	0.87
Grade 7	0.87	0.88	0.87	0.84	0.86
Grade 8	0.88	0.87	0.85	0.83	0.88
Ave	**0.86**	**0.86**	**0.84**	**0.82**	**0.85**
Min	**0.84**	**0.81**	**0.81**	**0.78**	**0.82**
Max	**0.88**	**0.88**	**0.87**	**0.84**	**0.88**

Figures 1 -3 – Scatterplots showing the relationship between state assessment scores and MAP in reading for three state tests at three grades.







Proficient Cut Score Estimates

For group 1, a proficient cut score estimate was generated using three methods, linear regression, second order regression, and Rasch Status on Standard. For group 2, a proficient cut score was generated using the distributional method.

Tables 4 and 5 show the proficiency cut score estimates for each subject in the five states using each method. In mathematics, the four methods produced highly similar estimates for the most part, with the range of estimates falling within three scale score points for all cases with one exception (grade 3 in Michigan). For reading, the difference in estimates generally ranged between one and four RIT with the exception of New Hampshire, where the difference in estimates ranged as high as seven RIT.

The tables also show the proportion of norm group students who performed at or below the estimated cut score, a metric which essentially reports the proportion of the NWEA norm group who would **not** have achieved the proficient score. In mathematics, the differences in the maximum and minimum estimated cut scores would produce differences of 0% to 9% in the proportion of students identified as proficient. For reading, these differences would produce differences of 3% to 9% in the proportion of students identified as proficient in four of the states. As expected, the range is far greater in New Hampshire, where the variance in cut score estimates produced differences 8% to 18% in the proportion of the norm population identified as proficient. This seems to have happened for two reasons. First, the linear and second order regression estimates for New Hampshire consistently lower estimates of the cut scores than the Rasch SOS and distributional methods. Second, because the New Hampshire standards for the upper

grades are near the middle of the NWEA norm distribution, the differences in scale score estimates created larger variation in the proportion of students identified below that score.

The tables do also show that the cut scores estimated for the five states in this study generally fell well below the 50th percentile in the NWEA norm distribution. This can affect the accuracy of some forms of estimation. For example, in a state that has set a low performance standard and uses a standard that may contain a floor effect artifact, linear regression estimates are very unlikely to produce a very accurate or repeatable estimate of the cut score.

Accuracy of Cut Score Estimates

We evaluated the accuracy of the cut score estimations by applying the results to the group 1 dataset. This allowed us to determine how well the cut scores predicted the actual performance of that group relative to proficiency. Because the linear regression, second-order regression and Rasch SOS estimates were generated from the group 1 data, the ability to generalize these cut scores to other populations may be limited. The distributional method, however, generated estimations of cut scores that came from a different population. Although many members of group 2 may also have been members of group 1, we would expect cut scores from the distributional method to generalize more effectively to a new population than the other estimates.

To evaluate the accuracy of cut score estimates from the linear regression, second-order regression and group 1 results, we applied the estimated cut scores to the same dataset. This method provides information about which cut score provides the best fit for the group being study, but does not generalize to the larger tested population in the state as well as the method employed with group 1.

Accuracy of the estimates was evaluated using three statistics. The first was correct prediction, that is, the percentage of cases in which the methodology correctly predicted the student's actual state test result. The second was the Type I error ratio, that is, the proportion of Type I errors among all errors in estimating proficiency on the state assessment. In general, the closer the Type I error proportion is to 50%, a situation in which the errors from underprediction and overprediction are equal, the more accurate the estimate of the cut score is assumed to be for that group. The third was the difference in the proportion of students MAP projected by that method to pass the state test and the proportion of students who actually passed the state examination. The closer the difference is to 0, the more accurate the estimate of the cut score for that group.

Table 4 – RIT Scale Proficiency Estimates for Reading/English Language Arts

	RIT Score Proficiency Estimate					Associated Percentile from NWEA Norms				
	Linear	Second Order	Rasch SOS	Distribution	Range	Linear	Second Order	Rasch SOS	Distribution	Range
Arizona										
Grade 3	189	191	190	190	2	24	28	26	26	4
Grade 4	196	198	198	200	4	23	27	27	32	9
Grade 5	201	204	203	204	3	21	27	25	27	6
Grade 6	207	209	208	208	2	25	30	27	27	5
Grade 7	208	211	210	211	3	21	27	25	27	6
Grade 8	213	216	215	216	3	24	31	28	31	7
Colorado										
Grade 3	190	190	189	191	2	26	26	24	28	4
Grade 4	202	202	201	202	1	37	37	34	37	3
Grade 5	205	205	204	206	2	30	30	27	32	5
Grade 6	211	212	211	211	1	34	37	34	34	3
Grade 7	217	217	216	215	2	43	43	40	37	6
Grade 8	220	221	220	218	3	41	44	41	36	8
Delaware										
Grade 6	204	205	202	205	3	20	21	16	21	5
Grade 7	205	205	204	208	4	17	17	15	21	6
Grade 8	209	209	206	210	4	17	17	13	18	5
Michigan										
Grade 3	170	170	170	174	4	10	10	10	14	4
Grade 4	185	185	186	188	3	16	16	17	20	4
Grade 5	194	195	193	196	3	18	19	16	21	5
Grade 6	197	198	197	200	3	15	16	15	19	4
Grade 7	204	204	204	207	3	20	20	20	25	5
Grade 8	210	211	208	212	4	24	26	20	28	8
New Hampshire										
Grade 3	181	181	185	184	4	25	25	33	31	8
Grade 4	193	190	197	196	7	29	24	39	36	15
Grade 5	199	201	205	205	6	27	32	43	43	16
Grade 6	207	207	211	211	4	34	34	46	46	12
Grade 7	209	210	215	215	6	30	32	46	46	16
Grade 8	214	216	220	220	6	33	39	51	51	18

Table 5 – RIT Scale Proficiency Estimates for Mathematics

	RIT Score Proficiency Estimate					Associated Percentile from NWEA Norms				
	Linear	Second Order	Rasch SOS	Distributional	Range	Linear	Second Order	Rasch SOS	Distributional	Range
Arizona										
Grade 3	194	195	194	195	1	25	27	25	27	2
Grade 4	201	202	201	203	2	23	26	23	28	5
Grade 5	210	211	210	210	1	28	31	28	28	3
Grade 6	217	219	218	218	2	33	38	35	35	5
Grade 7	220	222	221	220	2	30	34	32	30	4
Grade 8	228	230	230	228	2	36	40	40	36	4
Colorado										
Grade 3	196	196	196	196	0	30	30	30	30	0
Grade 4	204	205	203	205	2	31	34	28	34	6
Grade 5	213	213	212	214	2	35	35	33	38	5
Grade 6	222	223	222	223	1	45	47	45	47	2
Grade 7	233	233	233	233	0	59	59	59	59	0
Grade 8	239	239	238	238	1	60	60	58	58	2
Delaware										
Grade 3	190	191	190	193	3	16	18	16	22	6
Grade 4	200	201	199	201	2	21	23	19	23	4
Grade 5	205	207	206	207	2	18	22	20	22	4
Grade 6	213	215	213	215	2	25	29	25	29	4
Grade 7	220	222	221	223	3	30	34	32	36	6
Grade 8	225	228	227	228	3	31	36	34	36	5
Michigan										
Grade 3	172	171	168	174	6	5	4	3	7	4
Grade 4	186	186	185	188	3	9	9	8	12	4
Grade 5	200	200	199	201	2	19	19	17	21	4
Grade 6	210	211	208	209	3	29	32	25	27	7
Grade 7	216	218	216	217	2	30	35	30	32	5
Grade 8	219	221	221	221	2	26	30	30	30	4
New Hampshire										
Grade 3	187	188	190	189	3	32	35	41	38	9
Grade 4	198	198	199	199	1	32	32	35	35	3
Grade 5	205	206	208	207	3	31	34	40	37	9
Grade 6	216	216	216	216	0	44	44	44	44	0
Grade 7	221	223	224	224	3	42	47	49	49	7
Grade 8	230	231	231	231	1	50	53	53	53	3

In general, all methods produced highly accurate pass/fail predictions with no method falling below 81% correct in any state. On average, the rates of correct pass/fail prediction were almost identical across all methods, with an average range of 84.2% to 84.5% in reading/English language arts and an average range of 86.1% to 86.2% in mathematics. Thus rates of correct prediction were not useful in differentiating the accuracy with which the methods predicted the performance of group 1.

All methods employed in the study tended to produce greater numbers of Type I errors than Type II errors. In reading, the distributional method produced Type I error rates that were considerably lower than the other three methods. In mathematics, the distributional method also produced Type I error rates that were well below the other methods, although second order regression also produced Type I error rates below 60%.

Since all methods employed produced more errors of overprediction than underprediction, it would be expected that the estimates would tend to overstate estimated number of proficient students on the MAP test relative to actual performance on the state exams. The distributional method came the closest to producing RIT estimates the matched actual performance on the state examination, overestimating the pass percentage by an average of 3.0% in reading and by 2.1% in mathematics. This was interesting since the distributional method was the only one in which estimates generated with one sample were applied to a new sample. Of the other methods, second order regression produced lower rates of overestimation than linear regression and Rasch SOS, overestimating the pass percentage by an average of 4.7% in reading and by 2.2% in mathematics.

Table 6 – Accuracy of estimates in reading/English language arts

Method	Correct	Wrong	% Correct	Type I error	Type I error ratio	Difference in proficiency prediction & actual performance
Linear						
Arizona	12645	2409	84.0%	1793	74.4%	+8.8%
Colorado	12219	2689	82.0%	1537	57.2%	+2.1%
Delaware	5595	902	86.1%	660	73.2%	+6.4%
Michigan	12724	2218	85.2%	1582	71.3%	+6.3%
New Hampshire	4256	743	85.1%	578	77.3%	+8.3%
Unweighted Averages			84.5%		70.7%	(average absolute difference) 6.4%
Second Order						
Arizona	12728	2326	84.5%	1404	60.4%	+3.8%
Colorado	12213	2695	81.9%	1468	54.5%	+0.9%
Delaware	5592	905	86.1%	642	70.9%	+5.8%
Michigan	12719	2223	85.1%	1520	68.4%	+5.5%
New Hampshire	4244	755	84.9%	560	74.2%	+7.3%
Unweighted Averages			84.5%		65.7%	(average absolute difference) 4.7%
Rasch SOS						
Arizona	12689	2365	84.3%	1622	68.6%	+5.6%
Colorado	12161	2747	81.6%	1728	62.9%	+4.0%
Delaware	5584	913	85.9%	752	82.4%	+9.1%
Michigan	12717	2225	85.1%	1627	73.1%	+6.9%
New Hampshire	4194	805	83.9%	330	41.0%	-2.9%
Unweighted Averages			84.2%		65.6%	(average absolute difference) 5.7%
Distributional						
Arizona	14574	2503	85.3%	1571	62.8%	+3.7%
Colorado	17084	3840	81.6%	2322	60.5%	+3.8%
Delaware	5576	921	85.8%	575	62.4%	+3.5%
Michigan	12701	2241	85.0%	1242	55.4%	+1.6%
New Hampshire	4198	801	84.0%	342	42.7%	-2.3%
Unweighted Averages			84.3%		56.8%	(average absolute difference) 3.0%

Table 7 – Accuracy of estimates in mathematics

Method	Correct	Wrong	% Correct	Type I error	Type I error ratio	Difference in proficiency prediction & actual performance
Linear						
Arizona	14712	2255	86.7%	1649	73.1%	+6.1%
Colorado	17027	2858	85.6%	1517	53.1%	+0.9%
Delaware	10879	1856	85.4%	1101	59.3%	+6.7%
Michigan	12923	1952	86.9%	1330	68.1%	+5.9%
New Hampshire	4309	699	86.0%	442	63.2%	+3.8%
Unweighted Averages			86.2%		63.4%	(average absolute difference) 4.9%
Second Order						
Arizona	14826	2141	87.4%	1329	62.1%	+3.0%
Colorado	17028	2857	85.6%	1417	49.6%	-0.1%
Delaware	10084	1832	84.6%	1178	64.3%	+2.7%
Michigan	12948	1927	87.0%	1194	62.0%	+3.1%
New Hampshire	4313	695	86.1%	390	56.1%	+1.8%
Unweighted Averages			86.2%		59.5%	(average absolute difference) 2.2%
Rasch SOS						
Arizona	14767	2200	87.0%	1522	69.2%	+5.0%
Colorado	16994	2894	85.5%	1696	58.7%	+2.5%
Delaware	10860	1875	85.3%	1282	68.4%	+5.4%
Michigan	12897	1978	86.7%	1429	72.2%	+5.9%
New Hampshire	4299	709	85.8%	331	46.7%	-0.9%
Unweighted Averages			86.1%		67.1%	(average absolute difference) 4.7%
Distributional						
Arizona	14735	2232	86.8%	1227	55.0%	+4.6%
Colorado	17031	2854	85.6%	1396	48.9%	-0.3%
Delaware	10878	1857	85.4%	1043	56.2%	+1.8%
Michigan	12932	1943	86.9%	1129	58.1%	+2.1%
New Hampshire	4305	703	86.0%	349	49.6%	0.0%
Unweighted Averages			86.2%		54.6%	(average absolute difference) 2.1%

Discussion

The approach employed in this study takes steps to address some of the issues commonly raised in regard to linking tests by Thissen (2005) and others. The assessment used to link state assessments in this study provides direct and immediate feedback on performance to students, parents, and educators and is used to inform instruction and for accountability by school systems. While these stakes are not identical to those that may exist on some state tests, students clearly have more incentive to perform their best on this assessment than NAEP. Because the content on the MAP assessment is aligned to state standards, the two linked assessments are more likely to both reflect the content that is expected to be taught in the curriculum than is possible with NAEP. This should reduce the error in prediction that could be

attributed to content differences. Finally the estimates for this study were either derived directly from populations who had taken both tests or from schools who were known to have tested almost all their students on both tests. This helped assure that variance in cut score estimation that could be attributed to differences in the MAP and state sample populations would be minimized.

This is not to criticize efforts to use of NAEP as one means to attempt to link results on the various state tests to one another. Our prior studies linking the MAP assessment to state proficiency cut scores have basically reinforced the findings from prior NAEP studies. In particular, our prior research has reinforced prior findings around the wide variance in the rigor of state proficiency cut scores.

In terms of this study, the Pearson correlations between the MAP mathematics assessments and the five companion state assessments were strong, indicating that MAP has good predictive validity for this group of tests. In addition, all methods employed to test cut score estimates predicted the actual pass/fail performance of students with a high rate of accuracy. The cut score estimates generated for mathematics by the distributional method produced proficiency rate estimates that were very close to those actually achieved by students and also produced the most desirable balance between type I and type II errors among the four methods.

While the evidence of predictive validity was not as consistent in the reading domain, the correlations between state-aligned versions of MAP and their companion state tests were consistently strong and the distributional method produced estimates of proficiency in reading across the five states tested that nearly equaled those in mathematics.

These effectiveness of the distributional method in predicting proficiency outcomes was a little surprising, because it was the only method in which we tested the accuracy of the estimate on a different population than the one used to generate the original estimate. Because our intention was to employ this method in a study to estimate cut scores in new states, the effectiveness of the distributional method in predicting results in this group of states was encouraging.

All of this would suggest that the MAP scale could be used to produce a reasonably refined ranking of the rigor of mathematics and reading standards for most states at a given point in time. For purposes of this study we were also interested in determining whether MAP cut score estimates could be used to accurately project the rate of proficiency that the NWEA norm population might demonstrate on a state assessment. In these five states, we found that the best cut score estimates derived from the MAP scale overstated the actual performance of students by two to three percent. While this is a unfortunate, it would not lead to the conclusion that cut scores estimated from MAP would be likely to greatly understate the rigor of state standards or that they could not be used to project the approximate performance of another group on a state assessment.

The objectives set within the No Child Left Behind Act are extremely ambitious, to have all students demonstrate proficiency in reading, mathematics, and science by 2014. The definition of what constitutes proficiency is left to states. This seems somewhat akin to saying that the standard of proficiency for

basketball players is being able to dunk the basketball, without specifying the height of the basket. Whether that's desirable is best left for another discussion. We do believe, however, that reporting proficiency rates without attempting any means of comparison across states leaves us in a position in which we know that many students are dunking the academic basketball but we don't know the height of the basket used.

As the name of the Act implies, adoption of the No Child Left Behind Act was driven by the principle that requiring proficiency will help assure that traditionally underserved minority students will have the same expectations and opportunities as other students by requiring they achieve the same results. Of course one can achieve equity and eliminate achievement gaps entirely by having all students demonstrate their ability to dunk on a three-foot basket, and one will also eliminate achievement gaps if students are asked to dunk on a fifteen-foot hoop. The point is that one can't know if they're achieving meaningful equity of opportunity through proficiency measures without knowing the difficulty of the standard being attempted.

This is why, as schools work to achieve this goal and as policy makers consider making changes to the Act, it becomes more important to have a better understanding of what proficiency really means in the fifty states. Efforts to establish the difficulty of these standards relative to a common scale can show us the degree to which states have attempted to reflect the intent of the law in their standards. They can help policy-makers be more critical of states in which high rates of achievement may be, at least in part, a product of low proficiency standards. And they can help policy-makers be more supportive of states which may show lower rates of achievement but excellent progress toward truly rigorous standards.

The methodological challenges inherent in attempting to evaluate the various state proficiency challenges are large, but not daunting. Efforts to link these standards to common scales should continue and researchers should continue to investigate and implement innovations to improve these processes.

References

Angoff, W.H. (1964). Technical problems of obtaining equivalent scores on tests. *Journal of Educational Measurement, 1,* 11-13.

Braun, H. & Qian, J. (2005). *Mapping State Performance Standards on the NAEP Scale.* Educational Testing Service. Princeton, N.J.

Cronin, J. (2006, April). *The Effect of Test Stakes on Growth, Accuracy and Item-Response Time on a Computer-Adaptive Test.* Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco: CA.

Dorans, N & Holland, P.W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Meausrement, 37,* 281-306.

Feuer, M.J., Holland, P., Green, B.G. Bertenthal, M.W., & Hemphill, F. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Report of the Committee on Equivalency and Linkage of Educational Tests, National Resource Council. Washington, DC: National Academy Press.

Flanagan, J.C. (1964). Obtaining useful comparable scores for non-parallel tests and test batteries. *Journal of Educational Measurement, 1*, 1-4.

Holland, P.W., & Dorans, N.J. (2006). Linking and equating test scores. In R. Brennan (Ed.) *Educational Measurement* (4th ed.). Westport, CT: Praeger Publishers.

Inglebo (1997). *Probability in the Measure of Achievement.* Chicago: Mesa Press.

Kentucky Department of Education (1993). *Kentucky Instructional Results Information System 1991-92 technical report.* Frankfort, KY: Author.

Kingsbury, G.G. (2003, April). *A long-term study of the stability of item perameter estimates.* Paper presented at the annual meeting of the American Educational Research Association. Chicago: IL.

Kingsbury, G.G., Olson, A., Cronin, J., Hauser, C., & Houser, R. (2003, Nov.) *The state of state standards: Research investigating proficiency levels in fourteen states.* Northwest Evaluation Association. Lake Oswego, OR.

Koretz, D. (2005). *Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian.* Paper presented at the ETS Conference, Linking and Aligning Scores and Scales: A Conference in Honor of Ledyard R. Tucker's Approach to Theory and Practice. Princeton, NJ: Educational Testing Service.

Lennon, R. T. (1964). Equating non-parallel tests. *Journal of Educational Measurement, 1,* 15-18.

Lindquist, E.F. (1964) Equating scores on non-parallel tests. *Journal of Educational Measurement, 1,* 5-9.

Linn, R.L. (2005, December). *Adjusting for Differences in Tests.* Paper pressed for a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs. The Board on Testing and Assessment - The National Academies: Washington, DC.

Linn, R.L. (2005). *Test-based Educational Accountability in th Era of No Child Left Behind. Center for the Study of Evaluation Report 651.* National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

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

McGlaughlin, D. & Bandeira de Mello, V. (2003, June). *Comparing state reading and math performance standards using NAEP.* National Conference on Large-Scale Assessment, San Antonio, TX.

Mislevy, R.J. (1992, December). *Inking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Thissen, D. (2005, May). *Linking Assessments Based on Aggregate Reporting: Background and Issues.* Paper presented at the conference on Linking and Aligning Scores and Scales. Educational Testing Service. Princeton, NJ.