American Institutes for Research®

Vertically Moderated Standards: Logic, Procedures, and Likely Classification Accuracy of Judgmentally Articulated Performance Standards

Steve Ferrara
Eugene Johnson
Wen-Hung (Lee) Chen

American Educational Research Association

2004 Annual Meeting

April 12–16, San Diego, Calif.

Vertically Moderated Standards: Logic, Procedures, and Likely Classification Accuracy of Judgmentally Articulated Performance Standards

Steve Ferrara, Eugene Johnson, and Wen-Hung (Lee) Chen

American Institutes for Research

April 13, 2004

Paper presented in H. Huynh (Organizer), *Vertically Moderated Standards: Assumptions, Case Studies, and Applications*, and invited symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego.

Comments and questions to the authors at sferrara@air.org, ejohnson@air.org, lchen@air.org.

Vertically Moderated Standards: Logic, Procedures, and Likely Classification Accuracy of Judgmentally Articulated Performance Standards

Steve Ferrara, Eugene Johnson, and Wen-Hung (Lee) Chen
American Institutes for Research
April 13, 2004

Psychometricians have been searching for decades for statistical procedures to link tests developed from different specifications—that is, linking procedures that produce stable and validly interpretable results (e.g., Ercikan, 1997; Johnson, 1998; Johnson, Cohen, Chen, Jiang, & Zhang, 2003; Linn & Kiplinger, 1995; McLaughlin, 2003; National Research Council, 1998; Slinde & Linn, 1977). Most of this work has involved applying existing statistical equating procedures to link test *scores* from tests at the same grade level. Some of it has addressed linking test scores vertically, across grade levels and schooling levels (e.g., elementary to middle schools). Vertical scaling and linking of test scores has been most successful when test design and item selection within and across grade levels are managed carefully, so that sufficient overlap of items in adjacent test levels enables stable links. Examples of successful overlapping designs include K-12 commercial norm-referenced tests and individual intelligence and achievement tests. In some K-12 educational assessment situations, sufficiently overlapped test designs are not feasible.

The advent of NCLB requirements for tracking cohort growth and achievement gaps across grade levels has spurred new thinking. Some psychometricians have proposed and conducted socially moderated (i.e., judgmental) procedures to link test *performance standards* from two or more adjacent grade levels (e.g., Ferrara, 2003; Lissitz & Huynh, 2003). In this paper we describe (a) arguments that favor judgmentally linked performance standards over statistically linked test scores for some situations, (b) a judgmental approach for vertical linking of performance standards that has been used in an operational statewide assessment program, and (c) estimates of classification accuracy for vertically moderated performance standards, using generated data.

Only since the early 1990s have distinctions among types of test score links (i.e., equating, projection, statistical moderation, and social moderation) been explicated fully (see Linn, 1993; Mislevy, 1992). These explications provide a conceptual framework that has enabled creative thinking about procedures for linking scores and standards from different tests from the same grade level (e.g., Johnson, 1998; Johnson, Cohen, Chen, Jiang, & Zhang, 2003), and from tests for adjacent grade levels based on overlapping design specifications. The American Institutes for Research (AIR) works with a state assessment program and its advisory committees to design and develop an articulated assessment system for grades kindergarten through 8. This assessment system is intended to (a) meet No Child Left Behind requirements for assessing status within grade and annual growth across grades, and (b) achieve state goals to improve student achievement starting at kindergarten and continuing through middle and high school. We refer to this K-8 assessment system as ***articulated***, because within-grade content standards and performance standards specify proficient performance within the grade level, where proficient performance at a grade level predicts that students are on track to achieve proficient performance at the next adjacent grade.

The key concept in the vertical articulation process for setting performance standards is setting a performance standard in one grade that predicts performance in the subsequent adjacent grade. Specifically, and using grades 2 and 3 reading as an example, standard setting panelists considered the question, “What level of reading performance must students demonstrate in grade 2 in order to be considered on track for achieving proficient performance at grade 3?” Panelists considered this question with knowledge of the location of the grade 3 cut score projected onto the grade 2 score scale. A critical question about vertically articulated standards is: How accurately do they predict performance in subsequent grades?

A full study of classification accuracy would involve administering successive grade-appropriate versions of tests to a single cohort of students over a period of years and calculating the accuracy of classification. This process would require several years: Students would be given the grade 2 version of the test as second graders and would be classified. A year later, these same students, now in the third grade, would be given the grade 3 version of the test. The classification accuracy measure would be the degree to which prediction based on second grade classification on the second grade test matches the actual classification of the students on the third grade test, when they were in third grade. Of course, we need to know now, not later, whether the vertically articulated performance standards are accurate predictors of future performance. To speed up the evaluation, we have estimated classification accuracy of vertically articulated standards using simulated data generated under various assumptions of the changes in performance of the student population from grade 2 to grade 3.

The Vertical Articulation Process

Background

AIR is working with a state assessment program and its advisory committees to develop assessments in several content areas. The content area standards and assessments for grades K-8 are articulated in two ways. First, content standards overlap. At the highest level, reading content standards (e.g., apply strategies to comprehend and interpret literary, Informational, Technical, and Persuasive Text) are the same for grades K-12. Some grade-specific reading benchmarks appear at both grades 2 and 3 (e.g., establish a purpose for reading, make predictions, draw conclusions from text). Second, content area instruction in one grade builds on the previous grade's instruction, with the intention of preparing students to succeed in the subsequent grade. However, grade-level indicators of benchmarks and standards are grade-specific and Rasch scaling of items and examinees is conducted independently for grades 2 and 3.

The grade 2 reading diagnostic assessment contains 45 items. Approximately 25 are multiple-choice items; approximately 20 are open-ended. Maximum point values for the rubrics for these items range from 1 to 4. Of the 45 total items, 24 are discrete and 21 are associated with three reading passages. The grade 3 reading achievement test contains 36 items—29 multiple-choice, 4 short-response (scored 0-2), and 3 extended-response items (scored 0-4)—associated with four reading passages.

In this study we worked with the grade 2 reading diagnostic assessment and grade 3 reading achievement assessment. The state assessment program includes other diagnostic assessments below grade 2 and other diagnostic and achievement assessments above grade 3. The grade 3-8 assessments are in line with No Child Left Behind requirements. Standard setting procedures described here for reading in grade 2 and 3 were applied for reading assessments in other grades and for several grades in writing and mathematics.

Vertical Articulation Concept

The process for articulating performance standards across grades rests on the vertically articulated content standards. The concept of vertically articulated performance standards rests on the target performance standards: the Proficient standard on the grade 3 reading achievement assessment and the On Track standard on the grade 2 diagnostic assessment. The goal is that all students will perform at the Proficient standard or higher at grade 3. The purpose of the diagnostic assessment is to identify 2nd graders who, based on their performance on the grade 2 diagnostic assessment, are On Track in grade 2 to achieve at the Proficient level in grade 3. Schools and teachers then would have information about (a) which students are expected to achieve at the Proficient level on the grade 3 assessment, assuming they continue on the current achievement trajectory, and (b) for which students to provide intense instructional support in order to change the current achievement trajectory and increase chances that they will reach the Proficient level on the grade 3 achievement assessment.

AIR and this state assessment program considered two approaches prior to settling on vertical articulation: linear interpolation and judgmental standard setting with vertical IRT scaling. These procedures represented early thinking, when the intention was to have Proficient performance standards at each grade and a cut score above and below that standard. They also represent alternative conceptualizations of what it means to articulate performance standards across grade levels.

Linear interpolation. We proposed to set standards using the Bookmark procedure for kindergarten and grade 3, and then establishing cut scores for grades 1 and 2 using linear interpolation. Linear interpolation requires an assumption that growth in achievement across grades follows a straight-line trajectory with equal amounts of growth in achievement from one grade to the next. The steps in that process would have involved: (a) Identifying the percentile in the kindergarten empirical theta distribution corresponding to the kindergarten On Track cut score; (b) identifying the percentile in the grade 3 empirical theta distribution corresponding to the grade 3 Proficient cut score; (c) aligning the percentile scales for grades k, 1, 2, 3, and 4; (d) drawing a line across percentile scales from the kindergarten percentile to the grade 3 percentile; and (e) finding the percentiles in grades 1 and 2 that intersect that line. The state department of education and its technical advisory committee rejected this approach because of concerns about accepting the assumption of linear growth in achievement. This procedure is similar to that followed for the South Carolina Palmetto Achievement Challenge Tests (PACT; see Huynh, Meyer, & Barton, 2000).

Judgmental standard setting with quasi-vertical scale scores. We also proposed to set standards in each grade using the Bookmark procedure and assigning recommended cut scores to fixed Rasch scale scores. We would have placed all scores on the same apparent scale, across grades, as a convenience for reporting and interpretation, even though the individual grade scales would be completely independent of each other. The cross-grade metric would be set judgmentally. For example, the grade 3 Proficient cut score, determined in the Bookmark process, would be fixed at 350. Similarly, the grade 2 Proficient cut score on the independent grade 2 scale, also determined in the Bookmark process, would be fixed at 250. The cut scores above and below Proficient on the grade 2 scale would have been fixed to 275 and 225, respectively, and would correspond to the Proficient cut scores on the adjacent grades above and below grade 2, as judged by the panelists in their evaluation of the adjacent standards. That is, using grade 2 as an example, the score of 250 would correspond to the cut score selected by the panelists as identifying the level of attainment of a just barely Proficient 2nd grader. The 275 would correspond to the cut score on the grade 2 test that panelists felt would be achieved by a 2nd grader who was performing, on the second grade test, at the third grade Proficient level. The 225 would correspond to the score, again on the second grade test, of a 2nd grade student who was performing just barely at the 1st grade Proficient level. This procedure does not require assumptions about equal interval growth, as the cut scores are set separately for each grade. However, since the scores are set separately for each grade, there is no real guarantee that the various cuts would really be predictive of performance in the subsequent year. The state department of education and its technical advisory committee rejected this approach because of its conceptual complexity and concerns about the actual predictability across grades.

Standard Setting and Vertical Articulation

Unlike so-called vertical equating, the vertical articulation process does not involve statistical linking of scores across grades levels. It relies on the judgments of content experts about item responses requirements, the state reading content standards, and the performance level descriptors used for reporting test performance and standard setting.

Steps in the standard setting and articulation process. An AIR standard setting team trained 24 educators and community representatives on the state grade 3 reading content standards, assessment design, and performance level descriptors and on the Bookmark process. Nineteen of the panelists were teachers, three were other educators (e.g., a principal), and two were community people. Panelists learned about and then practiced the Bookmark process. Working in groups of four or five, they examined each item in the ordered item booklet and answered two questions for each item:

1. What does a 3rd grader need to know and be able to do in order to respond successfully to this item?
2. What makes this item more difficult than all items that precede it (in the ordered item booklet)?

Answers to these questions were intended to prepare panelists to make the Bookmark judgment for setting a cut score: “Place the bookmark on the page where you would expect two thirds of 3rd grade students who are *just barely Proficient* to respond successfully.” They also learned an alternate interpretation of the judgmental task: “Place the bookmark on the page where 3rd grade students who are *just barely Proficient* would have a 67% percent chance of responding successfully.” Panelists were trained to understand that students who are just barely Proficient would have less than a 67% chance of responding successfully to the item on the subsequent page and a 67% chance or higher on the previous items. In rounds 2 and 3 of the standard setting process, panelists received feedback on (a) pages on which other panelists placed their bookmarks (referred to as “agreement” information), and (b) percentages of students who would have reached the Proficient level (referred to as “impact” information). Panelists followed discussion procedures and used focus questions to assure that they examined and considered all feedback systematically. They were directed to consider the feedback information to clarify their thinking about item response requirements and difficulty and to reconsider the appropriateness the location of their bookmark in the ordered item booklet. After completing three rounds to establish the Proficient cut score, panelists followed similar procedures to establish an Advanced cut score above the Proficient cut score and a Basic cut score below, resulting in four performance levels in grade 3 reading. In all cases, the final cut score was determined as the theta score corresponding to the median page number, where the median page number was calculated across all panelists.

A second panel convened to establish On Track cut scores for each of grades kindergarten, 1, and 2. This panel of 24 included 19 teachers from grades k-3, three other educators (e.g., a local superintendent), a professor of reading, and a parent. They participated in training and practice as described above and, beginning with the grade 2 diagnostic assessment, examined the ordered item booklet and answered the two questions described above. Their training included specific focus creating articulated performance standards, the articulated design of the state reading content standards, and the procedures they would follow to articulate performance standards across grades. In round 1, panelists placed their bookmarks in the ordered item booklet in response to the direction, "Place the bookmark on the page where you would expect two thirds of 3rd grade students who are *just barely On Track* to respond successfully." In round 2, they received and discussed agreement information and impact information and discussed it systematically, as described above.

Then panelists received articulation feedback information and began the process of articulating the grade 2 On Track standard with the grade 2 Proficient standard. At this point a representative group of five members of the grade 3 standard setting panel participated with the K-3 panel in discussion of articulation feedback information. The articulation feedback information identified the page in the grade 2 ordered item booklet that corresponded to the same percentile score in the grade 2 scale score distribution as the percentile score in the grade 3 distribution that corresponded to the grade 3 Proficient cut score. Panelists were instructed to consider the item and item response requirements (a) on this page, (b) the page on which their individual bookmarks were currently located, and (c) the pages in between. They were instructed to consider whether they should reconsider their judgments about item response requirements and placement of the bookmark in light of this new information. Specifically, panelists considered whether students who are just barely On Track should be expected to respond successfully to any or all of the items between their current grade 2 book-marked page and the projected grade 3 page, and whether 67% of those students who are *just barely On Track* should be expected to respond successfully to an item in that sequence of pages. In round 3 panelists again considered agreement, impact, and articulation feedback.

Panelists followed the same procedures to articulate On Track performance standards for grade 2, grade 1, and kindergarten. Other standard setting panels followed similar procedures to establish vertically articulated standards in two other content areas. In all cases, the final cut score was determined as the theta score corresponding to the median page number, where the median page number was calculated across all panelists.

Vertically Articulated Standards

The standard setting panels recommended final cut scores of 1.27 in grade 2 and 0.88 in grade 3. The K-2 standard setting panel intentionally set the grade 2 cut score higher (in the grade 2 theta metric) than the grade 3 cut score (in the grade 3 metric). These panelists made clear in discussion that they were more concerned about grade 2 students who need remediation not receiving that remediation, and less concerned about some grade 3 students who reached On Track in grade 2 receiving additional remediation. In these discussions they were referring in everyday terms to false negative errors (in the first case) and false positive errors (in the second case).

We present final recommended cut scores (i.e., page numbers in the ordered item booklet) for the grade 3 reading proficient level and the grade 2 On Track level in Table 1. We also have included impact and articulation information for grades K-2.

Table 1. Grade 3 Proficient Cut Score and On Track Cut Scores for Grades K-2 and Accompanying Feedback

		Impact						
Page Number	**Percentage Raw Score**	**All Students**	**Female**	**Male**	**White**	**Black**	**Hispanic**	**Multi-racial**
Grade 3 Proficient cut score								
24	61	72	76	70	78	48	59	67
Grade 2 articulation information								
45	-- [1]	72	75	69	77	51	61	67
Grade 2 On Track cut score								
45	75	72	76	69	77	51	62	67
Grade 1 articulation information								
48	-- [1]	70	73	66	73	53	55	70
Grade 1 On Track cut score								
47	73	77	80	75	80	64	66	77
Grade 1 articulation information								
49	-- [1]	77	81	74	79	67	67	72
Grade 1 On Track cut score								
54	82	68	72	64	70	56	59	61

Notes. Maximum possible score and last page number is 49 in grade 3, 65 in grade 2, 62 in grade 1, and 65 in kindergarten. “Impact” is percentage of students in each group achieving the Proficient or On Track level.
[1] Panelists did not receive this information.

As is evident in Table 1, the page number that corresponds to the final panel cut scores corresponds closely to the articulation page number for grades 2 and 1. Likewise, the percentage of all students who would have achieved the On Track level in grades 2 and 1 are similar to the percentage reaching Proficient in grade 3. The page number that corresponds to the final panel cut score for kindergarten is higher than the articulation page number. As might be expected, the percentage of all students reaching the On Track level is lower in kindergarten than in the other grades. Table 2 provides information on the influence of the three types of feedback—agreement, impact, and articulation—on bookmark placements in round 2 of standard setting.

Table 2. Influence of Feedback Information on Bookmarked Page Numbers in Round 2

	Panel	**5 Panelist Tables**		**24 Panelists**	
	Median	**Lowest**	**Highest**	**Lowest**	**Highest**
Grade 2					
Round 1	45	43	49	40	49
Round 2	45	44	48	43	49
Grade 1					
Round 1	47	44	55	43	56
Round 2	47	47	55	46	55
Kindergarten					
Round 1	52	46	58	41	58
Round 2	52	51	56	50	58

Table 2 indicates that the overall panel's median bookmarked page did not change as a result of discussion at the beginning of round 2 of the feedback information from round 1. However, feedback and discussion appears to have influenced individual table medians and individual panelist bookmarked pages. In round 2, tables and individual panelists with bookmark placements earliest in the ordered item booklet moved their bookmarks closer to the table and panel median page numbers. Because the three types of feedback were presented at the beginning of round 2, it is not possible to distinguish the influence of articulation information on bookmark placement decisions in round 2. Panelist responses to the workshop evaluation form provide some insight. Of the 19 K-2 panelists who completed an evaluation form:

- 12 strongly agreed that the articulation information gave them information they needed to complete their assignment (6 agreed with the statement).

- 10 strongly reported that the articulation information was very important in their placement of the bookmark (7 reported that it was somewhat important, 2 reported that it was not important); the corresponding numbers were 11, 7, and 1 for agreement data and 8, 9, and 2 for impact data.

Sixteen of 18 responding panelists reported general satisfaction with the placements of the three On Track cut scores. One panelist would have moved the kindergarten bookmark one page lower. One panelist would have moved the grade 2 bookmark two pages higher; four would have moved the grade 1 bookmark an average of over 6 pages higher.

Panelist discussions at the beginning of rounds 2 and 3 during standard setting for grades K-2 tended to focus on two general topical areas: (a) Whether all students and On Track students at each grade could be expected to have learned the knowledge and skills required by the test items, and (b) setting fair performance standards. Panelists regularly referred to setting performance standards that are fair to students. They discussed fairness in two ways: Setting On Track cut scores that would identify (a) students in each grade who clearly would need the intensive intervention instruction that would ensue by not reaching the On Track cut score in order to reach On Track or Proficient levels in the subsequent grade, and (b) percentages of students for whom the range of school systems in this state could be expected to provide intensive intervention. In early in discussions panelists also discussed, in everyday logic terms, notions of weighing false positive and false negative rates against one another.

Standard setting panelists, the state department of education, and its advisory committees seemed satisfied with the recommended cut scores, impact information, and standard setting process. The State Board of Education adopted the cut scores recommended by the panels. Panelist comments suggested that they also were aware that the On Track performance standards may or may not prove in the future to be accurate predictors of reaching On Track and Proficient levels. A full study of classification accuracy would involve administering successive grade-appropriate versions of tests to a single cohort of students over a period of years and calculating the accuracy of classification. Of course, the state department of education needs to know now, not later, whether the vertically articulated performance standards are accurate predictors of future performance. In addition, as one or two panelists observed early in training, the goal is to intervene with students to assure that fewer and fewer students each year would not reach the On Track level in grades K-2 and that students who failed to reach On Track in a grade would reach On Track or Proficient in the subsequent grade. Instructional interventions would prevent accurate estimates of the classification accuracy of the On Track standards. To speed up the evaluation process, we have generated data to evaluate the accuracy of the On Track grade 2 cut score for predicting grade 3 reading Proficient performance. We describe the process of simulating data and discuss classification accuracy results in the next section of this paper.

Simulation Study

In order to evaluate the likely classification accuracy of the grade 2 On Track cut score, we generated data for three *types* of growth (i.e., growth models) and four *amounts* of growth. We estimated hypothetical distributions of grade 3 reading proficiency under three growth models:

- **Linear growth model**, in which the proficiency of all examinees increases by a fixed amount. Examinee positions in the distribution do not change relative to one another. This model serves as a benchmark for considering results from the other two growth models.

- **Remediation model**, in which the reading proficiency of examinees *below* the On Track level at grade 2 increases at grade 3 more than the proficiency of other examinees. This model reflects the possible outcome of intense remediation in reading during grade 3 for all students who did not reach the On Track level in grade 2.

- **"Rich Get Richer" model**, in which the reading proficiency of examinees *above* the On Track Level at grade 2 increases at grade 3 more than the proficiency of other examinees. This model reflects the possible outcome of no or ineffective remediation in reading during grade 3 for students who did not reach the On Track level in grade 2: Reading proficiency would increase more rapidly for students above the On Track level in grade 2 than for students who were below the On Track level in grade 2.

In addition, we examined four amounts of growth:

- **Negative growth.** All grade 3 thetas are .39 units lower than the empirical grade 2 thetas. This covers the distance between the grade 2 On Track cut score (i.e., 1.27 on the grade 2 theta scale) and the grade 3 Proficient score (i.e., .88 on the grade 3 theta scale). In this situation, the percentages of student achieving On Track at grade 2 and Proficient in grade 3 are equal.

- **No growth**. All grade 3 thetas are equal to the original grade 2 thetas. In this situation, the percentage of students reaching Proficient in grade 3 is higher than the percentage that reached On Track in grade 2. This is because the grade 3 cut score is lower in the grade 3 theta scale than is the grade 2 cut score in the grade 2 theta scale.

- **Low growth**. All grade 3 thetas are .39 units higher than the original grade 2 thetas.

- **Moderate growth**. All grade 3 thetas are .78 units higher than the original grade 2 thetas.

Method and Procedures

We started with the empirical distribution of grade 2 proficiency estimates (i.e., theta estimates from the grade 2 reading test) and estimated from that observed proficiency distribution hypothetical grade 3 proficiency distribution estimates for the same population of students for 12 conditions: four amounts of growth under three growth models. We estimated individual proficiencies using:

$$\theta_3 = \theta_2 + [\alpha + (\beta * f(\theta_2)],$$

where θ_3=the proficiency estimate at grade 3, θ_2=the proficiency estimate at grade 2, and α and β are defined as follows.

The parameter α has four conditions, defined in the four amount-of-growth conditions described above. The parameter β has three conditions, corresponding to the conceptual growth models described above. They are:

(1) β=0,

(2) β defined as $f\,(\theta_2) = (\theta_c - \theta_2)_+$, and

(3) β defined as $f\,(\theta_2) = (\theta_2 - \theta_c)_+$.

In conditions 2 and 3, the subscript “$_+$ “ indicates that the value of the quantity in the parentheses is returned if the quantity is positive, while a value of 0 is returned if the quantity is negative.

Results

Simulated theta distributions. Table 3 contains the descriptive statistics for the 12 type x amount growth conditions.

Table 3. Descriptive Statistics for the Grade 2 Empirical Data and 12 Sets of Simulated Data

Growth Amount	Mean	SD	Max	Min	Skewness	Kurtosis
Grade 2 empirical data						
Grade 2 empirical	1.77	1.11	6.42	-3.79	-0.34	1.13
Linear growth model (N=9,933)						
Negative	1.38	1.11	6.03	-4.18	-0.34	1.13
No	1.77	1.11	6.42	-3.79	-0.34	1.13
Low	2.16	1.11	6.81	-3.40	-0.34	1.13
Moderate	2.55	1.11	7.20	-3.01	-0.34	1.13
Remediation growth model (N=9,933)						
Negative	1.44	1.01	6.03	-2.92	0.05	0.75
No	1.83	1.01	3.42	-2.53	0.05	0.70
Low	2.22	1.01	6.81	-2.14	0.05	0.75
Moderate	2.61	1.01	7.20	-1.75	0.05	0.75
Rich Get Richer growth model (N=9,933)						
Negative	1.57	1.29	7.31	-4.18	-.03	0.81
No	1.96	1.29	7.70	-3.79	-.03	0.81
Low	2.35	1.29	8.09	-3.40	-.03	0.81
Moderate	2.74	1.29	8.48	3.01	-.03	0.81

The means and standard deviations in Table 3 are as expected: Negative growth means are lower than no growth means, low and moderate growth means are higher than no growth means, standard deviations for the grade 2 empirical data and the linear and remediation growth means remain unchanged, the standard deviations for the remediation growth model are slightly reduced by the reduction in skewness and shift to the right introduced by the growth model, and the standard deviation for the Rich Get Richer growth model increase somewhat, also by the reduction in skewness and shift to the right introduced by the growth model. The grade 2 empirical mean and standard deviation reflect the effects of re-centering of the full grade 2 item bank on the items in the ordered item booklet used in standard setting.

Line graphs representing each of the growth type x amount models appear in Figures 1-3.

Figure 1 . Distributions of Emipirial and Simulated Thetas Under Linear Growth Model with Different Growth Sizes

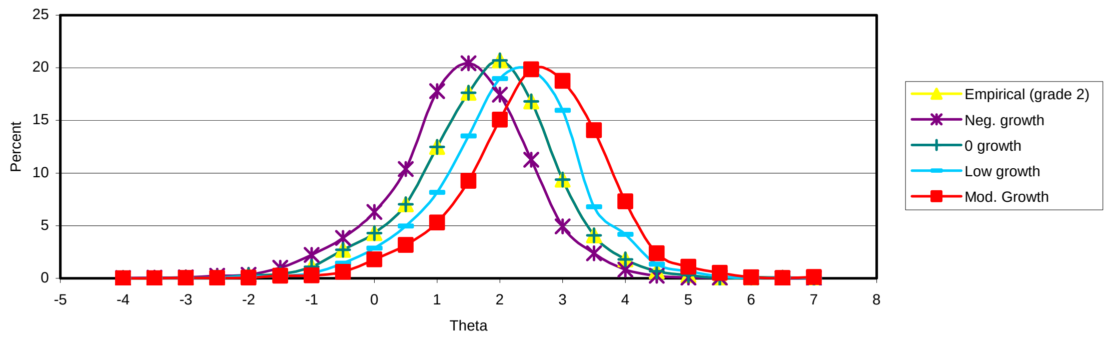


Figure 2. Distributions of Emipirial and Simulated Thetas Under Remediation Growth Model with Different Growth Sizes

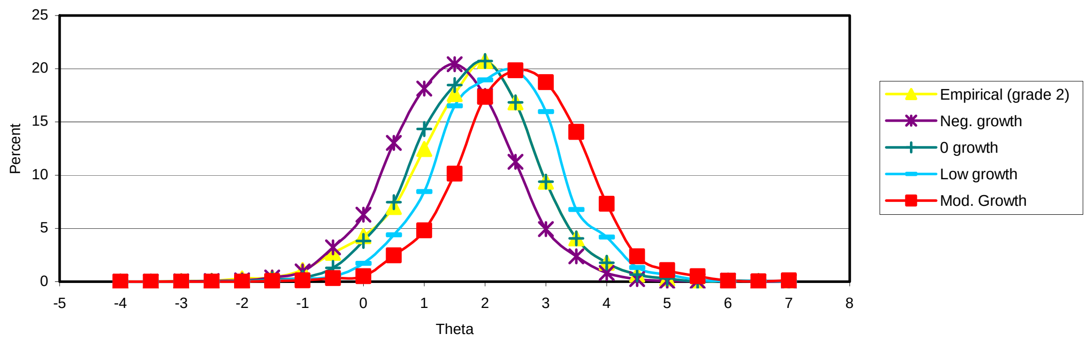
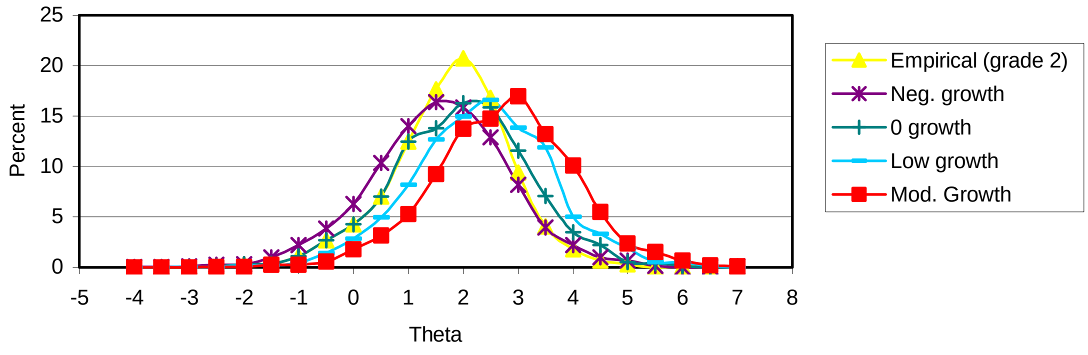


Figure 3 . Distributions of Emipirial and Simulated Thetas Under "Rich Get Richer" Growth Model with Different Growth Sizes



As with the descriptive statistics, the shapes and locations of these distributions are as expected. While the effects of the remediation growth model on the shapes of the line graphs in Figure 2 may not be readily apparent to the eye, the effects of the Rich Get Richer growth model are apparent in Figure 3. For example, the low growth and moderate growth curves each have a “bump-out” to the right at θ=2.16 (i.e., the distribution mean of 1.77 plus .39 growth amount) and 2.55 (i.e., the distribution mean of 1.77 plus .78 growth amount) respectively, corresponding to the change in growth trajectories for those models.

Simulated grade 3 results. Table 4 contains the projected percentages of 3rd graders who would reach the Proficient level on the grade 3 reading assessment, based on applying the grade 3 cut score (θ=0.88) to the theta distributions estimated under the 21 growth model conditions.

Table 4. Percentages of Examinees Who Would Reach the Proficient Level in Grade 3 for Three Hypothetical Growth Models

Growth Amount	Percentage
Linear and Rich Get Richer growth models	
Negative	70.8
No	81.1
Low	87.5
Moderate	92.1
Remediation growth model	
Negative	70.8
No	83.6
Low	90.8
Moderate	95.1

As expected, percentages of simulated examinees reaching the Proficient level increase in Table 4 as the amount of growth increases. Results for the Linear and Rich Get Richer growth models are the same. This is because the Rich Get Richer model is the Linear model, stretched out on its right side.

Classification accuracy results. Agreement between empirical classification of grade 2 examinees below and at/above the On Track level and projected classification of those examinees below and at/above Proficient on the grade 3 assessment appear in Table 5. Table 5 contains percentages of hits (i.e., correct classification above or below the cut score on both tests), false positive errors (i.e., students On Track on the grade 2 assessment and below Proficient on the grade 3 assessment), and false negative errors (i.e., students not On Track on the grade 2 assessment who reached Proficient on the grade 3 assessment).

Table 5. Classification Accuracy of Grade 2 On Track Standard for 12 Hypothetical Grade 3 Scenarios

		Classification Errors		
Growth Amount	**Hits**	**False Negative**	**False Positive**	**κ**
Linear growth model				
Negative	100.0	0.0	0.0	--
No	89.7	10.3	0.0	.72
Low	83.3	16.7	0.0	.51
Moderate	78.7	21.3	0.0	.34
Remediation growth model				
Negative	100.0	0.0	0.0	--
No	87.3	12.7	0.0	.65
Low	80.0	20.0	0.0	.40
Moderate	75.7	24.3	0.0	.22
Rich Get Richer growth model				
Negative	100.0	0.0	0.0	--
No	89.7	10.3	0.0	.72
Low	83.3	16.7	0.0	.51
Moderate	78.7	21.3	0.0	.34
Note. **κ** = Kappa coefficient.				

Table 5 contains several interesting results. First, there are no false positive errors for any growth type X amount model. This occurs because the grade 2 cut score is high within the grade 2 theta scale (i.e., 1.27) relative to the location of the grade 3 cut score in the grade 3 theta scale (i.e., 0.88). This result is consistent with panelist discussions about fairness of the cut scores (see above). In addition, the results for the Linear and Rich Get Richer growth models are the same. Because the grade 2 cut score is higher in its distribution relative to the grade 3 cut score in its distribution, under the Linear model all students who reach On Track in grade 2 will reach Proficient in Grade 3. The same holds for the Rich Get Richer model. On its left side, the Rich Get Richer model is the Linear model, stretched out on its right side, beginning at each of the growth-amount starting points, all of which are at or above the grade 3 cut score. Also, the hit rates under these 12 hypothetical scenarios appear fairly high, at least under the No growth model. In the no-growth scenario for all three growth types, the K-2 standard setting panel set the grade 2 performance standard low enough on the grade 2 test difficulty scale (i.e., and the grade 2 reading proficiency distribution) to avoid false negative errors and high enough so that the hit rate is near 90% for all three growth models. This result also is consistent with panelist discussions about fairness of the cut scores. Specifically, the true positive classification rate for all 12 scenarios is 70.8%. The range of true negative classifications in these data range from a low of 4.9% for the moderate growth-Remediation model to a high of 18.9% in the no growth-Linear and no growth-Rich Get Richer models. (The negative growth scenario is not discussed here because the grade 3 cut score is selected to assure 100% hit rates.) Finally, consistent with the hit rates, the Kappa coefficients suggest that classification accuracy is more accurate for the Linear and Rich Get Richer growth models. The explanation for finding no false positive errors applies here, as well.

Identification of students for remediation in grade 3. Finally, false negative rates increase with growth amounts, as expected. As the distribution of grade 3 reading proficiency moves to the right (relative to the distribution of grade 2 proficiency), more and more examinees reach the grade 3 Proficiency cut score, which remains fixed at 0.88 in the grade 3 theta scale. This highlights the goal of this state assessment program, and identifies a paradox: As regular instruction and intensive remediation becomes increasingly effective in improving the reading proficiency of 3rd graders, the grade 2 diagnostic test increasingly over-identifies students for remediation.

Discussion and Conclusions

In this paper we attempted to anticipate plausible scenarios for the classification accuracy of a grade 2 diagnostic reading assessment that is linked to a grade 3 reading achievement assessment using a vertical articulation process. We described the process of setting a grade 3

Proficient performance standard using the Bookmark standard setting method. Then we described the process of setting a grade 2 On Track performance standard and using vertical articulation information as feedback to panelists. This feedback was intended to help panelists provide a grade 2 performance standard that is vertically articulated with the grade 3 performance standard. The vertical articulation process is a specific example of using social moderation to link performance standards across tests from adjacent grades. We used vertical articulation to link performance standards from these tests, rather than a statistical equating process to link scores from these tests, because differences in (a) test content and item formats and largely non-overlapping content standards dictated against overlapping items on these tests, and (b) reading skills—that is, the reading constructs—dictate against including overlapping items in each test.

We simulated grade 3 data under three growth types (i.e., No growth, Remediation growth, and Rich-Get-Richer growth) and four growth amounts (i.e., negative, zero, low, moderate). We applied the grade 3 reading assessment cut score in each of these 12 scenarios and examined the accuracy of the grade 2 assessment in classifying examinees as On Track to reach Proficient on the grade 3 assessment. We found that, under these 12 hypothetical growth scenarios, the grade 2 assessment is unlikely to misclassify examinees as false positives. If this finding holds for the real situation in spring 2004 and beyond, it would be consistent with fairness goals for the grade 2 cut score discussed explicitly by the standard setting panel. Of course, negative and positive classification errors are interdependent. We found that false negative error rates are high in these growth scenarios, ranging from 10% to as high as 24%.

It is important to remember the underlying assumption in vertical articulation of performance standards: Identifying grade 2 students who are On Track to reach Proficient in grade 3 assumes that students are on an achievement trajectory to reach that performance standard. Maintaining that achievement trajectory requires at a minimum that students will be taught and will learn the reading content standards that are assessed on the grade 3 reading assessment.

Use of Grade 2 Test Results for Decision-Making

These results point to an important resource-use question: In the No growth scenario, and under all three growth types, the grade 2 assessment over-identifies 10-12% of 2nd graders for intensive remediation in grade 3. In the Moderate growth scenario, over-identification is 21-24%. It is easy to imagine the potential new costs and strains on existing school staff to provide intensive reading remediation for 1/10th to 1/5th of 3rd graders, without identifying the specifics of that remediation. In a school with 100 3rd graders in four classes of 25 students each, each classroom teacher would have to figure out how to manage standard instruction and classroom management and provide intensive remediation for two to four students. Or, a school system could provide a intensive remediation via pull-out services. Of course, that would require reallocating current school staff from other responsibilities or adding to the staff, both at a considerable cost.

This state could reduce the remediation burden by providing new intensive remediation to only a portion of the identified students. While this may make sense from a resource-allocation point of view, it increases risk of making false positive errors and involves some political risk. The state and local schools would have to explain to parents and teachers why some students who did not reach the On Track level in grade 2 will not receive the intensive remediation that others will receive. A solution could involve using additional reading proficiency information (e.g., teacher recommendations, additional individual reading assessment) to assign all students below the grade 2 On Track level to levels of different intensity of remediation services.

Implications for Vertical Moderation of Performance Standards

The vertical articulation process seems, to this point, to remain promising. It provides an alternative to creating overlapping test designs where such designs may be undesirable or insupportable. The grade 2 On Track reading standard appears to be linked reasonably well with the grade 3 Proficient reading standard, with the caveat that it may over-identify students for remediation. Results from these analyses in simulated data are consistent with the fairness goals that standard setting panelists articulated.

These results are not directly relevant to concerns about Annual Yearly Progress reporting requirements under No Child Left Behind. (States are required to report grade-to-grade achievement trends in grades 3-8, but not in earlier grades.) However, these results suggest that the vertical articulation approach can be considered for setting performance standards for the assessments in grades 3-8.

These results may be plausibly generalizable to the performance standards for kindergarten and grade 1 reading assessments in this state assessment program. A single standard setting panel articulated the reading standards for kindergarten through grade 2. Presumably, they applied the fairness concept consistently to their judgments about On Track standards for all three grades. Standard setting panels in two other content areas discussed similar logic for articulating standards, as well. It may be reasonable to expect to see the over-identification of students for remediation in these other grades and content areas that we found in these analyses. It is not clear whether conducting simulations and analyses in these other grades and content areas is a worthwhile investment at this point. This state department of education will receive the first wave of actual results of grade 3 reading performance later this year.

References

Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education, 10* (2), 145-159.

Ferrara, S. (2003, June). *Linking performance standards: Examples of judgmental approaches and possible applications to linking to NAEP*. In (A. Kolstad, Moderator), Linking state assessment results to NAEP using statistical and judgmental methods. Presentation at the National Conference on Large Scale Assessment, San Antonio, TX.

Huynh, H., Meyer, P., & Barton, K. (2000). *Technical documentation for the South Carolina PACT-1999 tests*. Available from the South Carolina Department of Education, Columbia, SC.

Johnson, E. G. (1998). *Linking the National Assessment of Educational Progress (NAEP) to the Third International Mathematics and Science Study (TIMSS): A technical report*. Report No. 98-499. Washington, DC: US Department of Education.

Johnson, E. G., Cohen, Chen, W-H., Jiang, T., & Zhang, Y. (2003). *2000 NAEP – 1999 TIMSS Linking Report.* Available from the U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6* (1), 83-102.

Linn, R. L., & Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education, 8* (2), 135-155.

Lissitz, R., & Huynh, H. (2003). *Vertical equating for the Arkansas ACTAAP assessments: Issues and solutions in determination of adequate yearly progress and school accountability*. Report submitted to the Arkansas Department of Education, Little Rock, AR.

McLaughlin, D., & Bandeira De Mello, V. (2003). *Comparing state reading and math performance standards using NAEP*. In (Andrew Kolstad, Moderator), Linking state assessment results to NAEP using statistical and judgmental methods. Presentation at the National Conference on Large Scale Assessment, San Antonio, TX.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

National Research Council. (1998). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement, 14* (1), 23-32.