

Chapter 9 Field Test IRT Scaling and Linking Analyses..... 5

Introduction 5

 Figure 1. Major Goals and Activities for Field Test Statistical Analysis 6

Horizontal Scaling: IRT Calibration for a Single Grade 6

Vertical Scaling: Linking Across Multiple Grades..... 7

 Figure 2. Summary of Field-Test Vertical Linking Item Configuration..... 10

Vertical Scale Linking Design. 10

IRT Preprocessing and Item Inclusion/Exclusion Criteria. 11

 Table 1. Number of Items by Type in the Vertical Linking Design. 12

 Table 2. Unique Number of CAT Items and Performance Tasks (PTs) Administered and the Survivorship for Vertical Scaling. 13

 Table 3. Summary of ELA/literacy and Mathematics Items by Purpose and Claim. 14

 Table 4. Summary of ELA/literacy by Type and Purpose. 15

 Table 5. Summary of Mathematics by Type and Purpose. 16

IRT Models and Software..... 16

 Figure 3. Sample ICC Plot for a Dichotomous Item Demonstrating Good Fit..... 18

 Figure 4. Sample ICC Plot for a Dichotomous Item Demonstrating Poor Fit..... 18

 Figure 5. Sample ICC Plot for a Polytomous Item Demonstrating Good Fit..... 19

 Figure 6. Sample ICC Plot for a Polytomous Item Demonstrating Poor Fit 19

Item Fit. 19

Vertical Linking Via Stocking-Lord..... 20

Evaluation of Vertical Anchor Item Stability..... 21

 Table 6. Example of STUIRT Linking Methods and Output. 21

Vertical Scale Evaluation. 21

Horizontal and Vertical Scaling Results 22

 Table 7. Summary of Classical Statistics by Purpose for ELA/literacy. 23

 Table 8. Summary of Classical Statistics by Purpose for Mathematics. 24

 Table 9. Summary of Item Parameter Estimates for Horizontal Calibration Step..... 24

 Figure 7. ELA/literacy Item Fit Chi-Square Plots (Vertical Scaling) 25

 Figure 8. Mathematics Item Fit Chi-Square Plots (Vertical Scaling) 26

 Table 10. Summary of Likelihood Ratio χ^2 Test Statistics by Grade and Content Area. 27

Table 11. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4. 28

 Figure 9. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 3 to 4 28

Table 12. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5. 29

 Figure 10. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 4 to 5..... 29

Table 13. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6. 30

 Figure 11. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 5 to 6..... 30

Table 14. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6. 31

 Figure 12. Comparison of ELA/literacy *a*- and *b*-parameter estimates for linking Grade 7 to 6 31

Table 15. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7. 32

 Figure 13. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 8 to 7..... 32

Table 16. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: High School to Grade 8..... 33

 Figure 14. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking High School to Grade 8..... 33

Table 17. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4. 34

 Figure 15. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 3 to 4..... 34

Table 18. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5. 35

 Figure 16. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 4 to 5..... 35

Table 19. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6. 36

 Figure 17. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 5 to 6..... 36

Table 20. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6. 37

 Figure 18. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 7 to 6..... 37

Table 21. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7. 38

 Figure 19. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 8 to 7..... 38

Table 22. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: High School to Grade 8..... 39

 Figure 20. Comparison of Mathematics a - and b -parameter estimates for Linking High School to Grade 8..... 39

Table 23. Vertical Linking Transformation Constants from the Stocking-Lord Procedure. 40

 Figure 21. Distribution of WRMSD for ELA/literacy (Vertical Linking Items)..... 41

 Figure 22. Distribution of WRMSD for Mathematics (Vertical Linking Items)..... 42

 Figure 23. ELA/literacy Cumulative Distributions of Student Ability across Grades 43

 Figure 24. Mathematics Cumulative Distributions of Student Ability across Grades..... 44

Table 24. Summary of Vertically Scaled Student Ability Estimates and Effect Size. 45

 Figure 25. ELA/literacy Student Ability Distributions Across Grades 3 to High School 46

 Figure 26. Mathematics Student Ability Distributions Across Grades 3 to High School..... 46

 Figure 27. Boxplots of Theta Estimates across Grade Level for ELA/literacy 47

 Figure 28. Boxplots of Theta Estimates Across Grade Level for Mathematics..... 48

 Figure 29. ELA/literacy Test Information by Score Level and Combined for Grades 3 to 6 49

 Figure 30. ELA/literacy Test Information by Score Level and Combined for Grades 7 to High School..... 50

 Figure 31. ELA/literacy Total Test Information for Grades 3 to High School 51

 Figure 32. Mathematics Test Information and Score Level and Combined for Grades 3 to 6 52

 Figure 33. Mathematics Test Information and Score Level and Combined for Grades 7 to High School..... 53

 Figure 34. Mathematics Total Test Information for Grades 3 to High School..... 54

 Figure 35. IRT Standard Error Plots for ELA/literacy Grades 3 to High School (HS) 54

 Figure 36. IRT Standard Error Plots for Mathematics Grades 3 to High School (HS)..... 55

Establishing the Minimum and Maximum Scale Score 56

 Table 25. Lowest and Highest Obtainable Theta Values and Resulting Theta Scale Summary. 56

Cross-validation of Vertical Linking Results 56

 Figure 37. Cross-validation of Vertical Linking Results Comparing cumulative frequency distributions of theta (EAP) for ELA and mathematics obtained from the CRESST cross-validation... 58

Calibration Step for Item Pool 59

 Table 26. Distribution of Student Observations per Item in the Field Test Pool. 59

 Table 27. Summary of IRT Item Parameter Estimates for the Field Test Item Pool. 60

Figure 38. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement Level Setting Sample) and the Item Pool Calibrations Step for ELA/literacy	61
Figure 39. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement Level Setting Sample) and the Item Pool Calibrations Step for Mathematics	62
Table 28. Distributions of ELA/literacy Theta Estimates and Conditional Standard Error of Measurement.....	63
Table 29. Distributions of Mathematics Theta Estimates and Conditional Standard Error of Measurement.....	63
References	64

Chapter 9 Field Test IRT Scaling and Linking Analyses

Introduction

The primary purposes of the Smarter Balanced assessments are to provide valid, reliable, and fair information concerning students' English Language Arts/literacy (ELA/literacy) and mathematics achievement with respect to the Common Core State Standards in grades 3 to 8 and high school. An important allied goal is to measure students' annual growth toward college and career readiness in grade 11 ELA/literacy and mathematics. For federal accountability purposes and potentially for state and local accountability systems, students' ELA/literacy and mathematics proficiency must also be reported. To meet these goals requires many technical characteristics to be demonstrated as evidence in support of validity. For instance, students must be measured on a common scale within a grade and content area. The methodology used to accomplish these varied goals is Item Response Theory (IRT). This chapter explains the methods used to construct the Smarter Balanced measurement scales using IRT. A description of the major Field Test scaling and linking activities in support of these goals are summarized in Figure 1.

As demonstrated by years of successful application in K-12 testing programs, IRT methods have the flexibility and strength to support the Smarter Balanced Consortium goals. IRT methods are ideally suited to the assessments and measurement goals of Smarter Balanced in both establishing a common scale and ongoing maintenance of the program, such as new item development and test equating and enabling computer adaptive testing (CAT) to be conducted (Wainer, 2000). Mixed-item-format tests, such as the Smarter Balanced assessments, that consist of dichotomous (selected-response) items, short answer responses, and performance tasks can be combined together and scaled concurrently (Ercikan, Schwarz, Julian, Burket, & Link, 1998; Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 2006). The purpose of the IRT horizontal calibration and scaling was to place items and ability estimates onto a common scale in a grade and content area. Since the Common Core State Standards (CCSS) were intended to be coherent and articulate across grade levels, they provided a foundation for developing Smarter Balanced assessments that support inferences concerning student change in achievement (i.e., growth). One approach to modeling student growth across grades is to report scores on a common vertical scale. For instance, comparing the means and standard deviations for scale scores across grades on the same scale is an intuitive approach for evaluating growth for a variety of test users. Vertical scales assume that increasing student proficiency is demonstrated across different levels of the assessment. For the CAT administration, vertical scaling permits items to be used across different grade levels if required. Another advantage of vertical scaling is that growth expectations concerning the establishment of achievement levels across grades can be inspected and ordered by standard setting panelists.

The IRT scaling for Smarter Balanced was performed in two steps. The first step was used to establish the horizontal and vertical scales that were used to set achievement levels. In the first step, items were initially scaled horizontally, where items in a single grade and content area were concurrently (i.e., simultaneously) calibrated. The vertical linking was accomplished using common items administered across grades (e.g., the same items given in 3rd and 4th grades) and then placing consecutive grades onto the vertical scale. In the second horizontal calibration step, the remaining, and much larger, item pool (containing noncommon items, each administered only to one grade) was scaled using the items from the first phase as linking/common items. Procedures associated with the IRT horizontal scaling are presented first. The horizontal scaling is followed by a discussion of assumptions, the methods used for vertical scaling, and the Field Test results. A cross validation of the vertical scaling is also briefly described. Next, the scale properties of selected NAEP and PISA items are presented, which were included to give further context to the establishment of the Smarter Balanced achievement levels concerning national and international comparisons.

Figure 1. Major Goals and Activities for Field Test Statistical Analysis

	Primary Goals	Major Analysis Activities
Phase 1 (Vertical Scaling)	<ul style="list-style-type: none"> Establish horizontal and vertical scales Analyze items and student “tests” that are scored on an expedited schedule to support achievement level setting Produce classical and IRT item statistics Provide provisional student proficiency estimates 	<ol style="list-style-type: none"> Performed classical item analysis and DIF analysis for Smarter Balanced vertical scaling items (on- and off-grade) in each grade/content area Calibrated all Smarter Balanced vertical scaling items (on- and lower-grade) at each grade/content Performed vertical scaling with Grade 6 as the pivot/base grade using embedded vertical scaling items from the lower-grade Estimated student proficiency Finalized recommendations for lowest and highest values of theta
NAEP/PISA Item Analysis	<ul style="list-style-type: none"> Provide IRT item parameters for embedded NAEP/PISA items on the Smarter Balanced vertical scales 	<ol style="list-style-type: none"> Calibrated all Smarter Balanced vertical scaling items and NAEP items Calibrated all Smarter Balanced vertical scaling items and PISA items (in high-school) Performed horizontal linking in the respective grades with on-grade Smarter Balanced items as linking items using their vertically-scaled item parameters Provided the resulting item parameters for NAEP/PISA items for use in standard setting
Phase 2 (Item Pool Calibration)	<ul style="list-style-type: none"> Provide classical and IRT item statistics for the remainder of the Field Test items on the Smarter Balanced scale 	<ol style="list-style-type: none"> Calibrated all Smarter Balanced on-grade items ($n \geq 500$) at each grade/content Performed horizontal linking in the respective grades with on-grade items as linking items with their vertically-scaled item parameters Provided IRT parameter estimates for the item pool

Horizontal Scaling: IRT Calibration for a Single Grade

Many K-12 programs scale, perform, and equate horizontally in the context of annual year-to-year assessments. For horizontal scaling in Smarter Balanced, methods using simultaneous, concurrent calibration of items were conducted at each content area/grade level. The calibration approach relied on a hybrid of the common items approach and the randomly equivalent groups linking approach. The “common items” approach requires that items and tasks partially overlap and are administered to different student samples. For the “equivalent groups” approach, the test material presented to different student samples is considered as comparably “on scale” by virtue of the

random equivalence of the groups. The random equivalence was implemented using the linear-on-the-fly test (LOFT) administration. Since neither type of linking method is guaranteed to work perfectly in practice, the linking design incorporated both types of approaches. This is done by assembling partially overlapping test content and randomly assigning them to students. The result is a design that is both reasonably efficient and well structured for IRT calibration. For further details concerning implementation of the data collection design and sampling, see Chapter 7, “Test Design and Field Test Design, Sampling, and Administration”, of the Technical Report.

The student response data consisted of the combined CAT and performance task (PT) components that were intended to measure the designated English language arts/literacy (ELA/literacy) or mathematics constructs as defined by the respective Field Test blueprints. The first step of the analysis was to create an item by student matrix reflecting item scores as well as missing information by design. For a given grade and content area, the dimension of this sparse data matrix was the total number of students times the total number of unique items (i.e., scorable units). Since each student only took a small subset of the available items, the remaining cells of the matrix represented items that were not administered. A provision in many IRT software programs is made for this “not-presented” or “not-reached” information necessary when multiple test forms are present. Students received a score that ranged from zero to the maximum permissible score level for the item administered. Using this sparse data matrix, a single grade-level concurrent calibration of all item data was performed. The procedures described here assumed that a unidimensional structure within each grade level is supported by the dimensionality analyses from the Pilot Test (see Chapter 6). Also based on the Pilot, the two-parameter logistic (2-PL) and generalized partial credit model (GPCM) models were chosen and implemented using the IRT program **PARSCALE** (Muraki & Bock, 2003). Chapter 6, on the Pilot Test, can be referenced for the results and decisions concerning of the IRT model comparison.

Vertical Scaling: Linking Across Multiple Grades

Determining whether students are making sufficient academic growth has received increased attention stemming from the No Child Left Behind (NCLB) Federal legislation. More recently, in the Race-to-the-Top legislation, there is a renewed emphasis on inferences concerning growth. These changes are intended to refocus instructional emphasis and facilitate inferences regarding change in academic achievement and readiness. Race-to-the-Top uses the Common Core State Standards (Common Core State Standards Initiative, 2010), which are articulated across grades and targeted at college- and career readiness. One method for evaluating change from one grade level to another is to develop a single common scale for use across multiple grade levels. Students are then ordered along the vertical scale implying that there is a progression of learning, primarily along a major or dominant dimension. The Common Core State Standards specifies across-grade-level articulation of content that is consistent with the general specifications for the construction of vertical scales. Another definition of growth embodied in the Common Core State Standards is learning progressions, which demonstrate how learning unfolds and characterize academic change at a finer level. As a result, there is increased interest in characterizing the amount of change that occurs for individual students or groups of students as they progress across grades. For these reasons, a continuous vertical scale reflecting growth was desired for the Smarter Balanced ELA/literacy and mathematics assessments ranging from grade 3 to 8 and high school. By contrast, in the NCLB legislation, there was no requirement for defining the relationships between content and performance standards across grades. In many instances, assessments and content standards were developed in a somewhat piecemeal fashion because the legislation was phased in over several years. For example, NCLB legislation began with a requirement (in reading and mathematics) in each of three grade spans (grades 3 to 5, 6 to 9, and 10 to 12). Subsequently, states were required to “fill in the gaps” and have assessments at grades 3 to 8. In many states, scale scores were established independently in each grade, which made inferences across grade levels more difficult. By contrast,

vertical scales permit scores on assessments administered at different grades to be reported using a common scale. The difference in scale scores for students from one grade to the next higher grade is used as an indicator of “growth” or change in achievement.

To conduct vertical scaling, common items are often administered to students at other grade levels than the targeted grade level for which they were developed, and primarily administered. Vertical scales assume that there is substantial overlap in the construct across grade levels. An assumption for test equating and interchangeable test scores is that test content and technical characteristics are parallel. Comparing students or groups of students who take parallel forms will generally be strongly supported, and most decisions of consequence for students and schools have depended on these types of cohort comparisons. Vertical scaling is not strictly equating. Holland and Dorans (2006) proposed a taxonomy consisting of three levels of linking that correspond to prediction, aligning, or equating. In this taxonomy, vertical scaling is a form of aligning in which tests have similar constructs and reliability but have different levels of difficulty and test taker populations. Overlap in content standards at adjacent grades may support the proposition that test forms for adjacent grades measure a common construct, but differences in the standards and psychometric properties of the test forms (e.g., test difficulty) imply that these forms are not parallel, so they may be linked but not equated. In the case in which scores are not parallel but a common proficiency is measured, linking can still occur (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999).

Assumptions and Interpretive Cautions Concerning Vertical Scales. The establishment of a vertical scale implies a) an increase in the difficulty of the assessments as the grade level increases, and b) a generally greater student proficiency in higher grades relative to lower grades. Accordingly, at the item level, it is assumed that students at a higher-grade level will generally have a higher probability of correctly answering an item than students at a lower grade level. This basic proposition must be substantiated to ensure the validity and plausibility of the vertical scale. With a sufficiently large and diverse sample of students, scale score means and other quantiles of the score distribution are expected to increase with grade level with a somewhat smooth pattern that is not erratic.

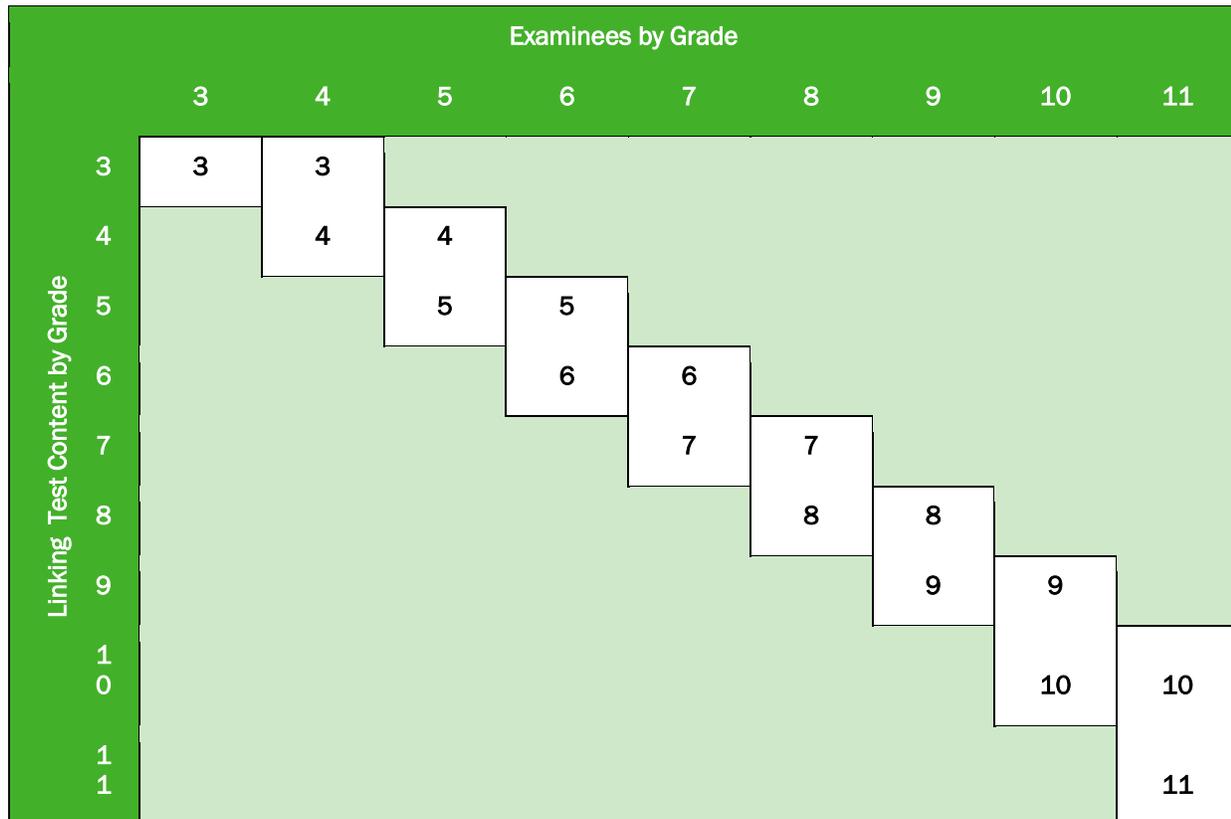
Validity evidence that vertical scales are appropriate for measuring students’ annual progress would include items that are easier in upper grades and have sensible patterns of variability within and across grades (i.e., meaningful separation of means and proficiency distributions across grades). The notion of interval level measurement dates back to taxonomy of measurement suggested by Stevens (1946). The familiar measures of height and weight, for example, exhibit this property. Interval level units are at best approximated in scales built for measuring latent variables such as academic ability. A ten-point difference in scale score units may mean something different at the low end of the score scale than it does at the middle or high score range. Achievement tests are constructed to have a strong first factor (i.e., essentially unidimensional), but multidimensionality to some extent will be reflected by changes in the content sampling across grades. The multidimensionality of the scale will be impacted by the relative importance of content subdomains at a given grade level and will determine the strength of these connections. Yen (2007) suggested that vertical scales are similar to a folding ruler that curves through space when held out. Connections among some levels of the scale are firmer, while others are somewhat looser. As a result, interpreting changes in scale scores is made more challenging when vertical scales are utilized. Additional validation efforts are appropriate when the change in scale scores over time is a focus of interest or for accountability purposes. In evaluating the vertical scale, an important question is whether the growth demonstrated by the vertical scale is consistent with expectations and the scope and sequence of grade-level instruction reflected in the Common Core State Standards. Growth needs to be interpreted in the context of the underlying scale variability. Finally, different vertical scaling methods can interact to stretch or compress the scale. Briggs and Weeks (2009), on the assessment they examined, suggested that the choice of the IRT scaling model had the largest impact on how growth

is depicted, followed by the choice of the calibration design and IRT scale score type (e.g., MLE, MAP, or EAP) chosen.

The IRT scaling and linking procedures described here were conducted in three successive stages. First, the scaling began with the evaluation of separate, horizontal (i.e., grade-specific), and concurrent scaling of items in the targeted item pool during Phase 1. After the grade-specific horizontal scaling was conducted in a content area, a separate, cross-grade vertical linking occurred using common items, also in Phase 1. The vertical scaling/linking was undertaken using the test characteristic curve transformation methods (Stocking & Lord, 1983). Using grade 6 as the baseline, each grade was successively linked onto the vertical scale separately for ELA/literacy and mathematics. Once the Smarter Balanced horizontal and vertical scales were established, the remaining items (i.e., the larger calibration item pool including the noncommon items) were linked horizontally onto this final scale in each grade and subject area in Phase 2. Another method for conducting vertical scaling is the multiple-group concurrent approach, which calibrates all grades simultaneously in a single step. The concurrent approach in vertical scaling context is a multigroup, nonequivalent group method that estimates underlying population distributions (latent means and standard deviations) for each group (Mislevy, 1987; Bock & Zimowski, 1997). Multigroup IRT permits the examination of group characteristics as a unit of analysis rather than as just individuals. For vertical scaling, the latent means should increase monotonically across grade levels. This method calibrates all students and grade levels in a single step.

Concurrent calibration uses all available item response information in the calibration and is therefore more efficient. Several studies that have investigated separate versus concurrent calibration have been inconclusive or limited in some respects, or found no substantive differences (Kim & Cohen, 1998; Hanson & Béguin, 1999; Ito, Sykes, & Yao, 2008). However, concurrent calibration can have limitations, such as convergence problems and restrictions on the number of items and observations the IRT software can handle. Kolen and Brennan (2004) suggested that separate linking steps might be preferable since it is more difficult to detect how items behave across grade levels or to diagnose any convergence problems in estimation, and violations of unidimensionality can be more problematic with the concurrent approach. The separate calibration approach produces two sets of item-parameter estimates, which can help identify and remediate potential problems. For example, if an item functioned poorly or was highly unstable across levels, it could be dropped as a vertical linking (common) item. This type of problem would be essentially undetectable with concurrent calibration, where all item parameters are estimated simultaneously, assuming common parameters across grade levels. Despite the utility of the multiple-group concurrent approach in other applications, and to reduce risk, concurrent calibration was not used as the vertical scaling method. The factors mentioned here and unknown effects of other factors such the size of the data matrix (item by student) across all grades, and the possibility of poor item functioning in the context of a Field Test led to this decision.

Figure 2. Summary of Field-Test Vertical Linking Item Configuration



Vertical Scale Linking Design. To implement the vertical scaling, representative sets of off-grade items were administered to an adjacent-upper grade. For example, grade 4 items were also administered to grade 5 students. To the extent possible, vertical linking item sets were intended to sample the construct that included both the CAT and performance task components and associated item types, and claims that conformed to the test blueprint. Linking items from the lower grade were administered to the upper-adjacent-grade level students, as shown in Figure 2. Content experts designated a target grade for each item and a minimum and maximum grade designation. Table 1 shows the vertical scaling linking design in terms of the number of CAT items and performance tasks. A set of six performance tasks was given on-grade, and the same set was administered off-grade for vertical linking. Each performance task had five or six items associated with it according to the test specifications. The same set of six performance tasks was administered in grades 9, 10, and 11 (high school). Table 2 presents the number of CAT items and performance tasks available for the vertical linking after test delivery and some item exclusions. In mathematics, particularly in grades 6 to 8 and high school (HS), a reduced number of items were available for vertical linking after test delivery, relative to the original test design. The total shown in Table 2 on the right is the number of items surviving after the IRT flagging criteria were applied, resulting in exclusion of some items (discussed in the next section). Other items might have been excluded in prior steps based on poor classical statistics. A full description of the item and test exclusion rules is given in Chapter 8 on “Field Test Datastep and Classical Test Analysis”. In some cases, a single item was eliminated from a given performance task. The resulting Smarter Balanced claim distributions for on-grade items and those targeted for vertical linking are presented in Table 3. In ELA/literacy, the claims are, respectively, Reading, Writing, Speaking/Listening, and Research, respectively. In mathematics, the

claims are Concepts and Procedures, Problem Solving, Communicating/Reasoning, and Modeling/Data Analysis. Tables 4 and 5 give a summary of item types by their purpose for ELA/literacy and mathematics.

IRT Preprocessing and Item Inclusion/Exclusion Criteria. Item functioning was evaluated prior to calibration and during the course of calibration in which items for which parameter estimates did not converge, or poorly functioning items, were excluded. In the data step, items were required to have 10 observations in a score category level for constructed-response (CR) items or 500 observations overall in order to obtain sufficiently stable IRT estimates. Many items, particularly in high school, were eliminated due to low numbers of observations. Chapter 8 has a complete description of the item- and student-exclusion rules applied and the resulting number of items available for vertical scaling. Some additional IRT-based rules are:

1. Local item dependence. ELA/literacy performance tasks contained a single “long-write” writing prompt that was subsequently scored for the dimensions of organization, elaboration, and conventions. These resulting scores were very highly correlated in the Field Test. Very high correlations between ratings of a single writing response can lead to local item dependence, which is a violation of IRT assumptions (Yen, 1993). As a result, the two dimensions for organization (0 to 4 score points) and elaboration (0 to 4 score points) were averaged and rounded for IRT scaling. This resulted in a score that ranged from zero to four points for the long-write performance tasks. In some cases, it was also necessary to collapse the top score because it had few or no observations.
2. Non-convergence results when item parameters could not be estimated in **PARSCALE**. Poor item parameter estimation was defined by either not achieving the criterion of largest estimate change lower than 0.005 or an erratic pattern of loglikelihood values. Standard errors were also evaluated along with item-parameter estimates as to their reasonableness.
3. For IRT analysis, all items with a -parameter estimates (i.e., discrimination) below 0.10 or the combination of a -parameter estimates below 0.20 and b -parameter estimates (i.e., difficulty) above 4.0 were excluded.
4. IRT parameter estimates, item characteristic curve plots, associated standard errors, along with item goodness-of-fit statistics were evaluated holistically to determine the quality of the resulting item parameter estimates. After examining these item characteristics, additional items were excluded due to poor functioning (e.g., a combination of very low discrimination and poor fit).
5. These criteria resulted in a subset of items in each grade being excluded due to poor IRT functioning. If a vertical linking item was excluded in the on-grade designation, it was also eliminated as a vertical linking item.

In general, after these exclusions were implemented overall convergence was met, and the resulting IRT item/ability parameter estimates under each model combination were reasonable.

Table 1. Number of Items by Type in the Vertical Linking Design.

Grade	CAT		PT		NAEP/PISA
	On-Grade	Off Grade	On-Grade	Off Grade	
ELA/literacy					
3	300	--	6		
4	300	150	6	6	75
5	300	150	6	6	
6	300	150	6	6	
7	300	150	6	6	
8	300	150	6	6	75
9		150		6	
10		150		6	75
HS	300	150	6	6	75
Mathematics					
3	300	--	6		
4	300	150	6	6	75
5	300	150	6	6	
6	300	150	6	6	
7	300	150	6	6	
8	300	150	6	6	75
9		150		6	
10		150		6	75
HS	300	150	6	6	75

Table 2. Unique Number of CAT Items and Performance Tasks (PTs) Administered and the Survivorship for Vertical Scaling.

Administered						Survivorship				
Grade	CAT		PT		NAEP/ PISA	CAT		PT		NAEP/ PISA
	On-grade	Off-grade	On-grade	Off-grade		On-grade	Off-grade	On-grade	Off-grade	
ELA/literacy										
3	306	-	6			261	-	6		
4	280	159	6	6	31	242	120	6	6	28
5	313	156	6	6		256	133	6	6	
6	292	160	6	6		232	131	6	6	
7	289	158	6	6		238	107	6	6	
8	300	161	6	6	30	243	123	6	6	30
HS	602	153	6	6	31/34	410	107	6	6	27/33
Mathematics										
3	320	-	6			304	-	6		
4	332	128	6	6	37	306	104	6	6	30
5	325	126	6	6		306	95	6	6	
6	327	127	6	6		222	102	6	6	
7	319	126	6	6		239	71	6	6	
8	333	128	6	6	36	230	73	6	6	33
HS	573	129	6	6	35/82	319	81	6	6	28/74

Table 3. Summary of ELA/literacy and Mathematics Items by Purpose and Claim.

Grade	Purpose	Number of Items	Claims (Percent)				
			1	2	3	4	Not Assigned*
ELA/literacy							
3	On-Grade	261	36	27	19	18	
	Off-grade	-	-	-	-	-	
4	On-Grade	242	30	28	21	21	
	Off-grade	120	33	28	22	17	
5	On-Grade	256	36	26	18	20	
	Off-grade	133	41	24	19	17	
6	On-Grade	232	31	29	19	21	
	Off-grade	131	40	24	19	17	
7	On-Grade	238	32	29	19	20	
	Off-grade	107	38	27	17	18	
8	On-Grade	243	34	27	20	19	
	Off-grade	123	40	28	20	13	
HS	On-Grade	410	44	31	10	16	
	Off-grade	107	45	23	20	12	
Mathematics							
3	On-Grade	304	61	6	15	6	12
	Off-grade	-	-	-	-	-	-
4	On-Grade	306	59	6	17	8	11
	Off-grade	104	56	4	15	7	18
5	On-Grade	306	59	6	16	7	12
	Off-grade	95	58	3	19	6	14
6	On-Grade	222	48	9	18	9	16
	Off-grade	102	59	4	13	7	18
7	On-Grade	239	56	4	16	9	15
	Off-grade	71	39	6	21	8	25
8	On-Grade	230	57	7	15	7	14
	Off-grade	73	49	4	12	10	25
HS	On-Grade	319	60	7	14	10	9
	Off-grade	81	57	6	15	7	15

Note: *Not Assigned refers to items that were not assigned to a claim at the time of the Field Test.

Table 4. Summary of ELA/literacy by Type and Purpose.

Item Purpose	Item Response Type	Score Type	Number of Items per Grade						
			3	4	5	6	7	8	HS
On-grade	SR*		115	92	95	81	77	91	142
	Other	Dichotomous	110	119	122	114	122	113	216
		Polytomous	36	31	39	37	39	39	52
Off-grade Vertical Linking Items	SR			57	53	46	27	38	39
	Other	Dichotomous		45	62	63	58	62	58
		Polytomous		18	18	22	22	23	10
NAEP	SR			22				20	12
	Other	Dichotomous		2				2	4
		Polytomous		4				8	11
PISA	SR								17
	Other	Dichotomous							12
		Polytomous							4

Note: *SR refers to selected-response.

Table 5. Summary of Mathematics by Type and Purpose.

Item Purpose	Item Response Type	Score Type	Number of Items per Grade						
			3	4	5	6	7	8	HS
On-grade	SR		48	65	78	21	39	41	66
	Other	Dichotomous	221	212	175	174	185	159	203
		Polytomous	35	29	53	27	15	30	50
Off-grade Vertical Linking Items	SR			11	12	31	9	7	18
	Other	Dichotomous		76	71	56	55	60	56
		Polytomous		17	12	15	7	6	7
NAEP	SR			20				19	18
	Other	Dichotomous		2				6	4
		Polytomous		8				8	6
PISA	SR								19
	Other	Dichotomous							44
		Polytomous							11

IRT Models and Software

Unidimensional IRT models were used to calibrate the selected-response and constructed-response (i.e., polytomous) items. Using the criteria and results from the Pilot Test and consultation with Smarter Balanced, the two-parameter logistic and the generalized partial credit models were chosen for use in the Field Test to establish the scale. For selected-response items, the two-parameter logistic (2PL) model was used (Birnbaum, 1968). The 2PL model is given by

$$P_i(\theta_j) = \exp[Da_i(\theta_j - b_i)] / \{1 + \exp[Da_i(\theta_j - b_i)]\},$$

where $P_i(\theta_j)$ is the probability of a correct response to item i by a test taker with ability θ_j ; a_i is the discrimination parameter; b_i is the difficulty parameter, for item i , and D is a constant that puts the θ ability scale into the same metric as the normal ogive model ($D=1.7$).

For constructed-response items, the generalized partial credit model (GPCM; Muraki, 1992) or partial credit model (PCM; Masters, 1982) is employed. The generalized partial credit model is given by

$$P_{ih}(\theta_j) = \frac{\exp \sum_{v=1}^h [Da_i(\theta_j - b_i + d_{iv})]}{\sum_{c=1}^{n_i} \exp \left[\sum_{v=1}^c Da_i(\theta_j - b_i + d_{iv}) \right]},$$

where $P_{ih}(\theta_j)$ is the probability of examinee j obtaining a score of h on item i , n_i is the number of item score categories, b_i is the item location parameter, d_{iv} is the category parameter for item i for category v , and D is a scaling constant given previously.

PARSCALE (Muraki & Bock, 2003) was used for the IRT calibrations. **PARSCALE** is a multipurpose program that implements a variety of IRT models associated with mixed-item formats and associated statistics. The psychometric properties of **PARSCALE** are well known, and it can efficiently and accurately calibrate large data sets such as those of Smarter Balanced assessments. The program implements marginal maximum likelihood (MML) estimation techniques for items and MLE estimation of theta.

The software program **STUIRT** (Kim & Lee, 2004) was used to conduct the vertical linking and horizontal linking in the item-pool calibration step. **STUIRT** implements four IRT scale transformation methods using the mean/sigma, mean/mean Haebara (1980) and Stocking-Lord (1983) methods. Consistent with previous research, the Stocking-Lord and Haebara methods are expected to have highly similar results (Hanson & Beguin, 2002). The Stocking-Lord transformation constants consisting of the slope (A) and intercept (B) terms were estimated and then applied to targeted item parameter estimates to place them onto the common vertical scale.

Since **PARSCALE** is limited in the types of graphical output, the program **PARPLOT** (ETS, 2009) was used to obtain item characteristic curves used for evaluating item functioning. A useful way to understand item functioning is to examine plots showing the observed and expected performance based on the item-parameter estimates. Figures 3, 4, 5, and 6 show example plots for a dichotomous and a polytomous item that demonstrate items with both good and poor fit. The solid line represents the expected item performance based on IRT, and the triangles represent the observed item performance, with the size of the triangles proportional to student sample size at a given level of theta. For an item to show good model data fit, it is expected that the triangles, especially the large-size ones, adhere closely around the item characteristic curves. Evaluation of item functioning was conducted visually using **PARPLOT** in conjunction with the goodness-of-fit statistic. Based on evaluation of the plots, any items demonstrating poor functioning were flagged and excluded from the calibrated item pool as previously described.

Figure 3. Sample ICC Plot for a Dichotomous Item Demonstrating Good Fit

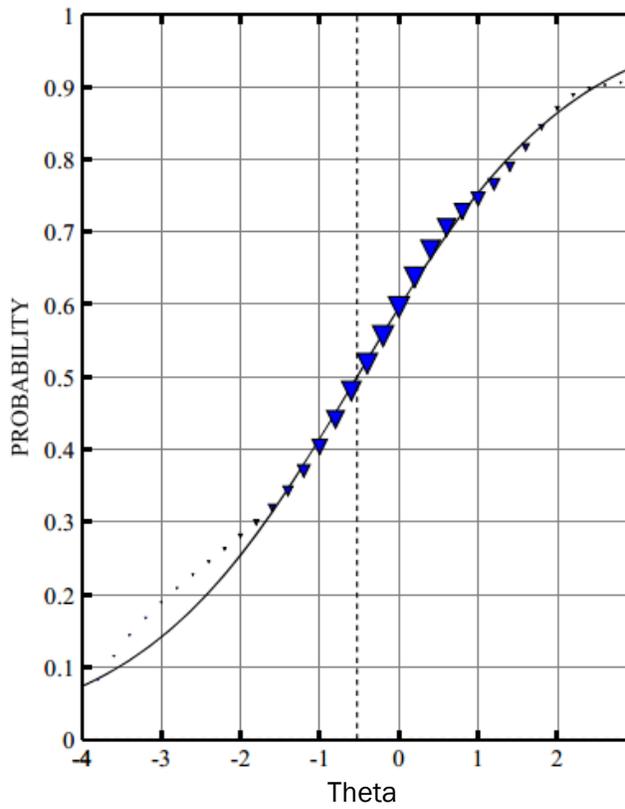


Figure 4. Sample ICC Plot for a Dichotomous Item Demonstrating Poor Fit

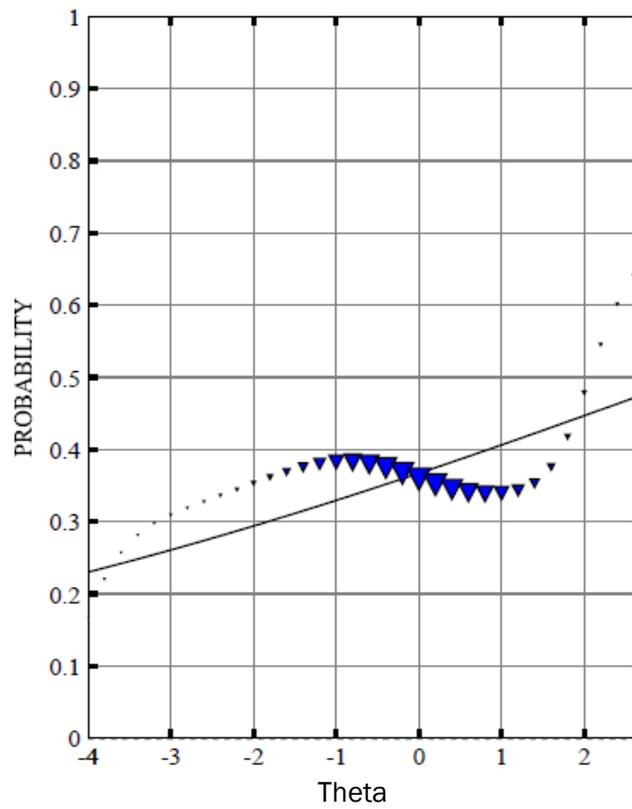


Figure 5. Sample ICC Plot for a Polytomous Item Demonstrating Good Fit

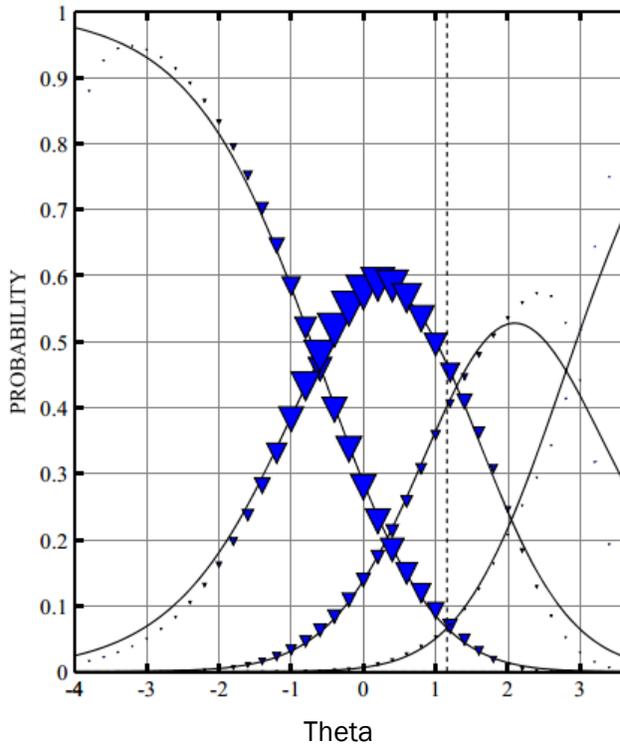
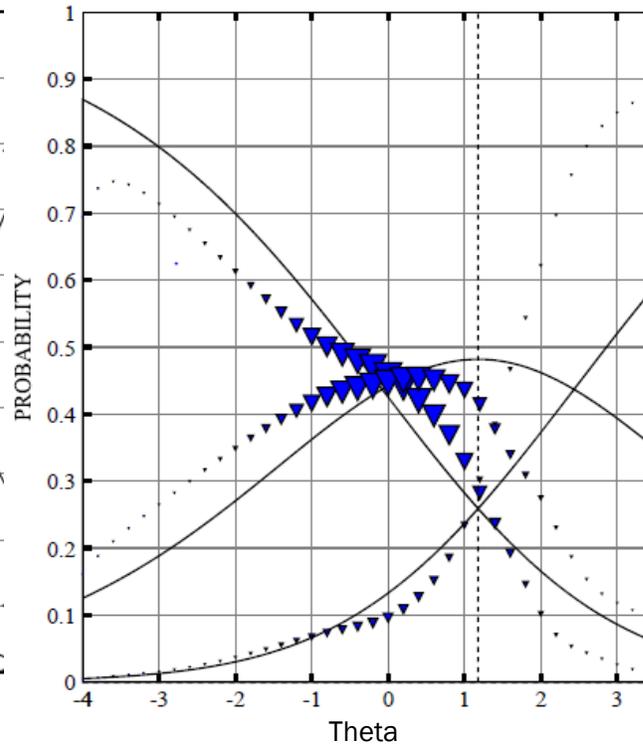


Figure 6. Sample ICC Plot for a Polytomous Item Demonstrating Poor Fit



Item Fit. The usefulness of IRT models is dependent on the extent to which they effectively reflect the data. Assessing fit in item response models usually involves validating assumptions underlying the models and evaluating goodness-of-fit, which specifically refers to how effectively the model describes the outcome data. IRT fit evaluation was conducted for calibrations using the two-parameter-logistic/generalized partial credit model (2PL/GPC) combination. The goodness-of-fit information contained in **PARSCALE** uses the likelihood ratio χ^2 test.

The likelihood ratio χ^2 test statistic can be used to compare the frequencies of correct and incorrect responses in the intervals on the θ continuum with those expected based on the fitted model (du Toit, 2003)

$$\chi_j^2 = 2 \sum_{h=1}^{n_g} \left[r_{hj} \log_e \left\{ \frac{r_{hj}}{N_h P_j(\bar{\theta}_h)} \right\} + (N_h - r_{hj}) \log_e \left\{ \frac{N_h - r_{hj}}{N_h [1 - P_j(\bar{\theta}_h)]} \right\} \right],$$

where n_g is the total number of intervals, r_{hj} is the observed frequency of correct responses to item j in interval h , N_h is the number of examinees in interval h , $\bar{\theta}_h$ is the average ability of examinees in interval h , and $P_j(\bar{\theta}_h)$ is the value of the fitted response function for item j at $\bar{\theta}_h$. The residuals are not under linear constraints, and there is no loss of degrees of freedom due to fitting the item parameters. The number of degrees of freedom is equal to the number of intervals remaining after neighboring intervals are merged, if necessary, to avoid expected values less than 5. Chi-square-type statistics tend to be sensitive to sample size (i.e., flagging more items with large sample size). Item

fit was evaluated in conjunction with other psychometric criteria and the plots described previously. No items were excluded based solely on fit.

Vertical Linking Via Stocking-Lord. The Stocking-Lord method was used as the primary method of linking adjacent grade levels to construct the vertical scale. In general, test-characteristic-curve methods such as the Stocking-Lord method have some advantages when compared to moment methods such as mean/mean or mean sigma (Baker & Al-Karni, 1991; Hanson & Béguin, 2002; Kolen & Brennan, 2004). When used with separate calibration, the test-characteristic-curve methods are more robust to violation of the IRT assumptions and produce less error when compared with moment methods. The Stocking-Lord procedure minimizes the sum of the squared differences over students between the target and reference test characteristic curves based on common items. Specifically, the procedure seeks to determine the slope (A) and intercept (B) that minimize the function

$$E = \frac{1}{N} \sum_{a=1}^N [T(\theta_a) - T^*(\theta_a)]^2,$$

where $T(\theta_a)$ is the test characteristic curve of linking items on the reference vertical scale and $T^*(\theta_a)$ is the test characteristic curve of linking items from the grade to be transformed onto the vertical scale. The linking takes place by applying the resulting slope and intercept to the targeted item parameters.

For 2-PL and GPC models, the transformations for discrimination and difficulty parameter estimates are

$$a^T = \frac{a}{A};$$

$$b^T = A \cdot b + B,$$

and for GPC model item-category parameters, the transformation is

$$d^T = A \cdot d.$$

The following transformations are applied to theta (ability)

$$\hat{\theta}^T = A \cdot \hat{\theta} + B.$$

The associated standard errors for the parameter estimates were transformed as follows

$$s.e.(\hat{\theta}^T) = A \cdot s.e.(\hat{\theta})$$

$$s.e.(\hat{b}^T) = A \cdot s.e.(\hat{b})$$

$$s.e.(\hat{a}^T) = s.e.(\hat{a}) / A$$

$$s.e.(\hat{d}^T) = A \cdot s.e.(\hat{d}).$$

The **STUIRT** program was used to implement the vertical linking using test-characteristic methods. An example of the **STUIRT** linking output comparing the different methods is shown in Table 6 for linking grade 3 ELA/literacy to grade 4. In this example, all methods produced similar slope and intercept

values. To implement the Stocking-Lord linking, the weights and quadrature points of the latent ability distribution output from **PARSCALE** were used. These quadrature points were transformed the same way as student abilities. The slope and intercept transformation parameters (A & B) were applied to the latent distributions produced by **PARSCALE** in each grade. **STUIRT** was also used to conduct the linking in Phase 2, where horizontal scaling of the remaining on-grade item pool was conducted.

Evaluation of Vertical Anchor Item Stability. An inspection of the differences between the off-grade estimates and the reference, on-grade ones for each vertical linking item was conducted. The weighted root mean squared difference (WRMSD) is calculated as

$$WRMSD = \sqrt{\sum_{j=1}^{N_g} w_j [P_n(\hat{\theta}_j) - P_r(\hat{\theta}_j)]^2},$$

where abilities are grouped in the intervals of 0.5 between -4.0 and 4.0 , $\hat{\theta}_j$ is the mean of the abilities in the interval j , N_g is the number of intervals, w_j is a weight equal to the proportion of estimated abilities from the transformed new form in interval j , $P_n(\hat{\theta}_j)$ is the probability of correct response based on the transformed new-item-parameter estimates at ability level $\hat{\theta}_j$, and $P_r(\hat{\theta}_j)$ is the probability of correct response at ability level $\hat{\theta}_j$ based on the reference-item-parameter estimates. A criterion of WRMSD greater than 0.125 was used to evaluate the linking. This criterion has produced reasonable results in other programs in year-to-year horizontal-equating contexts (Gu, Lall, Monfils, & Jiang, 2010). The distributions of WRMSD were evaluated; no linking items were eliminated based on the WRMSD statistic.

Table 6. Example of STUIRT Linking Methods and Output.

Method	Slope A	Intercept B
Mean/Mean	0.9627	-1.1864
Mean/Sigma	0.9585	-1.1823
Haebara	0.9533	-1.1683
Stocking-Lord	0.9444	-1.1889

Vertical Scale Evaluation. In the process of constructing the vertical scale, it was evaluated using a number of methods that included:

- correlation and plots of (common) item difficulties across grade levels;
- progression in test difficulty across grades;
- comparison of mean scale scores across grades;
- comparison of scale scores associated with proficiency levels across grades;
- comparison of overlap/separation of proficiency distributions across grades; and

- comparison of variability in scale scores (ability) within and across grades comparing scale score standard deviations.

Grade-to-grade change can be displayed as the differences between means and percentiles (10, 25, 50, 75, and 90) across grades. Separation of ability distributions can also be displayed by plotting the scale score cumulative distributions across grades. An index of separation in grade distributions suggested by Yen (1986) is the effect size. It standardizes the grade-to-grade difference in the means by the square root of the average of the within-grade variances. The effect size is defined as

$$\frac{\bar{\theta}_{higher} - \bar{\theta}_{lower}}{\sqrt{\frac{\sigma_{higher}^2 + \sigma_{lower}^2}{2}}},$$

where $\bar{\theta}_{higher}$ is the average ability estimate for the higher grade level, $\bar{\theta}_{lower}$ is the average ability estimate for the lower grade level, σ_{higher}^2 is the variance of the ability estimates for the higher grade, and σ_{lower}^2 is the variance of the ability estimates for the lower one.

Horizontal and Vertical Scaling Results

During classical item analysis, the performance of on-grade and vertical linking items was evaluated by comparing the item difficulty across the two adjacent grades. For a vertical sale to demonstrate change in achievement and be plausible, items are expected to be easier in the higher grade. For example, when an item is administered in grades 5 and 6, the p -value should be relatively higher (easier) in grade 6. The on- and off-grade average item difficulty and item-test correlations are given in Tables 7 and 8 for ELA/literacy and mathematics, respectively. Item difficulty is defined as the percentage of the maximum possible raw score. The average item difficulty is consistent with the notion of better performance in the higher grade-level necessary to establish a vertical scale. Since the tests differed widely in the number of items delivered, theta was used as the criterion rather than the typical total raw score for the item-test correlation.

The distributions (i.e., the five-number summary) for the IRT item discrimination and location parameters that resulted from the initial horizontal calibration in the vertical scaling are given in Table 9. The difficulty parameters (i.e., b -parameters) indicate that the tests were difficult, particularly in high school. Figures 7 and 8 present plots of chi-square item fit for ELA/literacy and mathematics. In general, there were only a relatively small number of outliers. Table 10 provides a summary of the χ^2 statistics and sample size ranges per item. Tables 11 to 22 show the distributions of untransformed item a and b -parameter estimates from common items for ELA/literacy and mathematics, respectively. Figures 9 to 14 for ELA/literacy and 15 to 20 for mathematics show plots of untransformed, vertical linking, item parameter estimates across grades. The item parameter estimates for the most part cluster along the diagonal. As a rule-of-thumb, the b -parameter estimate correlations should typically be above .90 and the a -parameter estimates above .85, which indicate the items behaved consistently across grades. The distributions of the untransformed a and b -parameter estimates are given for the common, vertical linking items across grades, and these parameters are plotted along with the correlations.

After the conclusion of the horizontal scaling and IRT item exclusion steps, the vertical scaling was conducted. Using grade 6 as the base and the common items across grade levels, each grade level was successively linked onto the vertical scale using the associated Stocking-Lord transformation constants. Table 23 presents the Stocking-Lord transformation constants that were obtained from **STUIRT**. For evaluative purposes, the WRMSD was computed, and histograms of the resulting values

were plotted in Figures 21 and 22 for each vertical linking set. To construct a criterion for evaluation, a vertical line plotted at 0.125 was used as a criterion for identifying items with large values. Consistent with the high correlations between linking items, the values WRMSD were for the most part below 0.10. This information was used diagnostically to evaluate the linking, and no items were removed based on the WRMSD.

The ability or theta estimates used were maximum likelihood estimates (MLEs) produced by **PARSCALE**. In cases in which an MLE could not be produced by **PARSCALE**, table driven sufficient statistics (Lord, 1980) were used to derive a theta estimate. Table 24 summarizes the resulting theta distribution for the ELA/literacy and mathematics vertical scales. It presents the five-number summaries, the means and standard deviations, and sample sizes along with the effect sizes. The effect size demonstrates the degree of change over grades and is not uniform with a larger change observed in the lower grades. Figures 23 and 24 display the cumulative distributions of ability (theta) for the vertical scale. The cumulative distributions of ability are more widely separated at the lower-grade levels, with diminishing amounts of change in the upper-grade levels in both ELA/literacy and mathematics.

Table 7. Summary of Classical Statistics by Purpose for ELA/literacy.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-grade	Number of Items	261	242	256	232	238	243	410
	Mean Difficulty	0.34	0.35	0.38	0.35	0.34	0.36	0.34
	Item-Total Correlation	0.51	0.50	0.52	0.49	0.49	0.49	0.47
Off-grade (Vertical Linking)	Number of Items		120	133	131	107	123	107
	Mean Difficulty		0.45	0.45	0.42	0.36	0.38	0.36
	Item-Total Correlation		0.54	0.52	0.52	0.51	0.51	0.49

Table 8. Summary of Classical Statistics by Purpose for Mathematics.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-grade	Number of Items	304	306	306	222	239	230	319
	Mean Difficulty	0.39	0.36	0.32	0.30	0.27	0.24	0.24
	Item-Total Correlation	0.59	0.58	0.56	0.60	0.59	0.53	0.53
Off-grade (Vertical Linking)	Number of Items		104	95	102	71	73	81
	Mean Difficulty		0.51	0.40	0.37	0.32	0.31	0.32
	Item-Total Correlation		0.62	0.61	0.58	0.62	0.59	0.56

Table 9. Summary of Item Parameter Estimates for Horizontal Calibration Step.

Grade	ELA/literacy							Mathematics						
	3	4	5	6	7	8	HS	3	4	5	6	7	8	HS
No. of Items	261	362	389	363	345	366	517	304	410	401	324	310	303	400
a-parameter														
Mean	0.598	0.594	0.599	0.580	0.577	0.577	0.546	0.756	0.765	0.752	0.736	0.814	0.749	0.727
SD	0.215	0.207	0.204	0.208	0.207	0.222	0.208	0.253	0.255	0.296	0.274	0.360	0.329	0.318
Min	0.117	0.171	0.138	0.165	0.164	0.119	0.129	0.149	0.206	0.200	0.126	0.103	0.146	0.126
10%	0.331	0.326	0.342	0.320	0.307	0.299	0.283	0.431	0.437	0.405	0.380	0.295	0.338	0.351
25%	0.430	0.451	0.451	0.433	0.429	0.419	0.405	0.563	0.577	0.537	0.555	0.513	0.477	0.478
Median	0.578	0.586	0.597	0.553	0.556	0.564	0.540	0.769	0.751	0.708	0.723	0.835	0.735	0.693
75%	0.723	0.736	0.735	0.719	0.723	0.722	0.675	0.941	0.952	0.936	0.923	1.103	0.971	0.929
90%	0.897	0.854	0.867	0.856	0.843	0.870	0.795	1.075	1.102	1.167	1.100	1.241	1.179	1.161
Max	1.186	1.317	1.222	1.320	1.249	1.434	1.349	1.379	1.457	1.816	1.587	2.053	1.831	1.885
b-parameter														
Mean	1.060	0.779	0.637	0.817	0.996	0.923	1.110	0.588	0.504	0.783	0.852	1.128	1.335	1.397
SD	1.256	1.295	1.187	1.251	1.220	1.364	1.358	1.261	1.134	1.021	1.190	1.140	1.205	1.172
Min	1.707	2.587	3.101	2.025	2.196	3.183	2.019	3.261	2.987	1.909	4.064	2.522	1.893	2.425
10%	0.528	0.757	0.879	0.814	0.478	0.721	0.577	0.985	1.017	0.518	0.794	0.278	0.172	0.103
25%	0.073	0.246	0.244	0.069	0.093	0.084	0.122	0.361	0.265	0.113	0.075	0.486	0.458	0.639
Median	0.971	0.671	0.618	0.753	0.969	0.925	0.977	0.701	0.552	0.831	0.923	1.187	1.390	1.445
75%	2.028	1.613	1.459	1.699	1.754	1.713	1.904	1.362	1.290	1.518	1.696	1.808	2.052	2.164
90%	2.773	2.552	2.247	2.347	2.569	2.772	2.946	2.208	1.956	2.074	2.297	2.228	2.952	2.790
Max	4.745	4.897	4.296	4.479	4.720	4.875	5.781	4.865	3.562	4.700	4.139	5.008	4.810	4.372

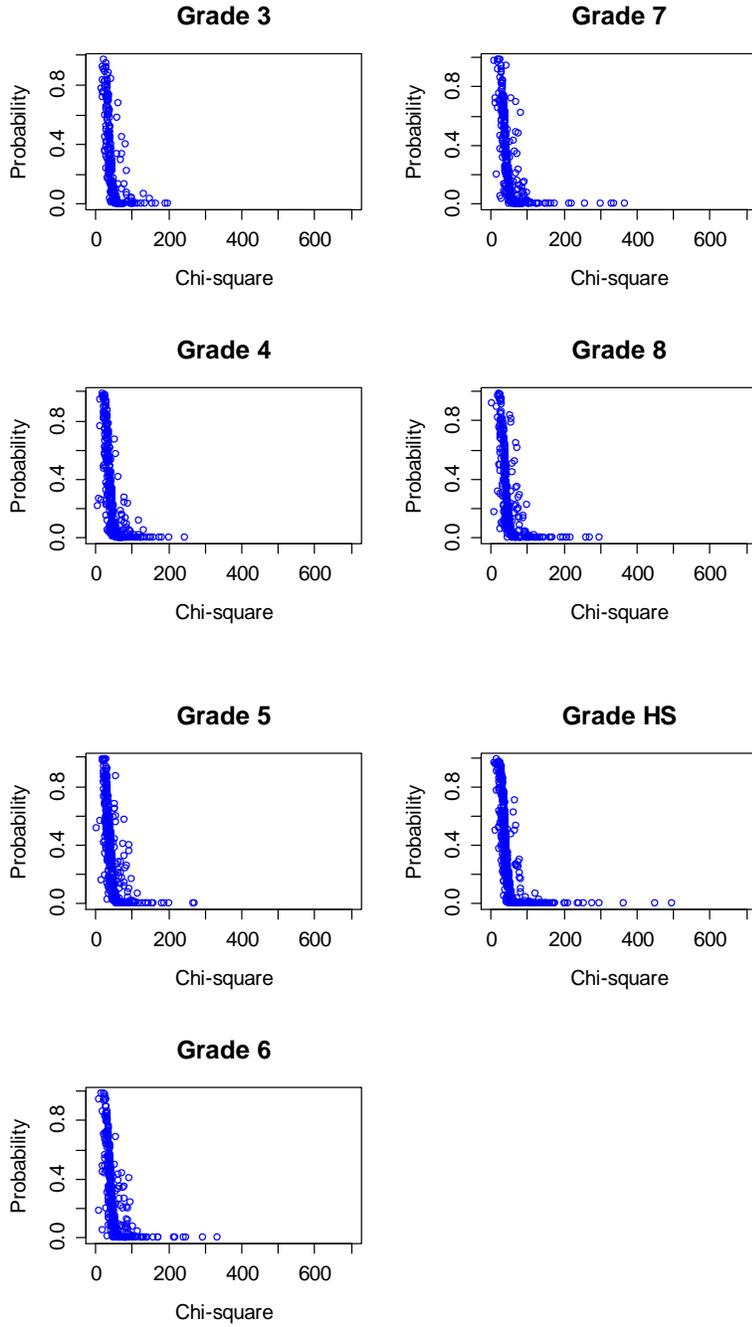


Figure 7. ELA/literacy Item Fit Chi-Square Plots (Vertical Scaling)

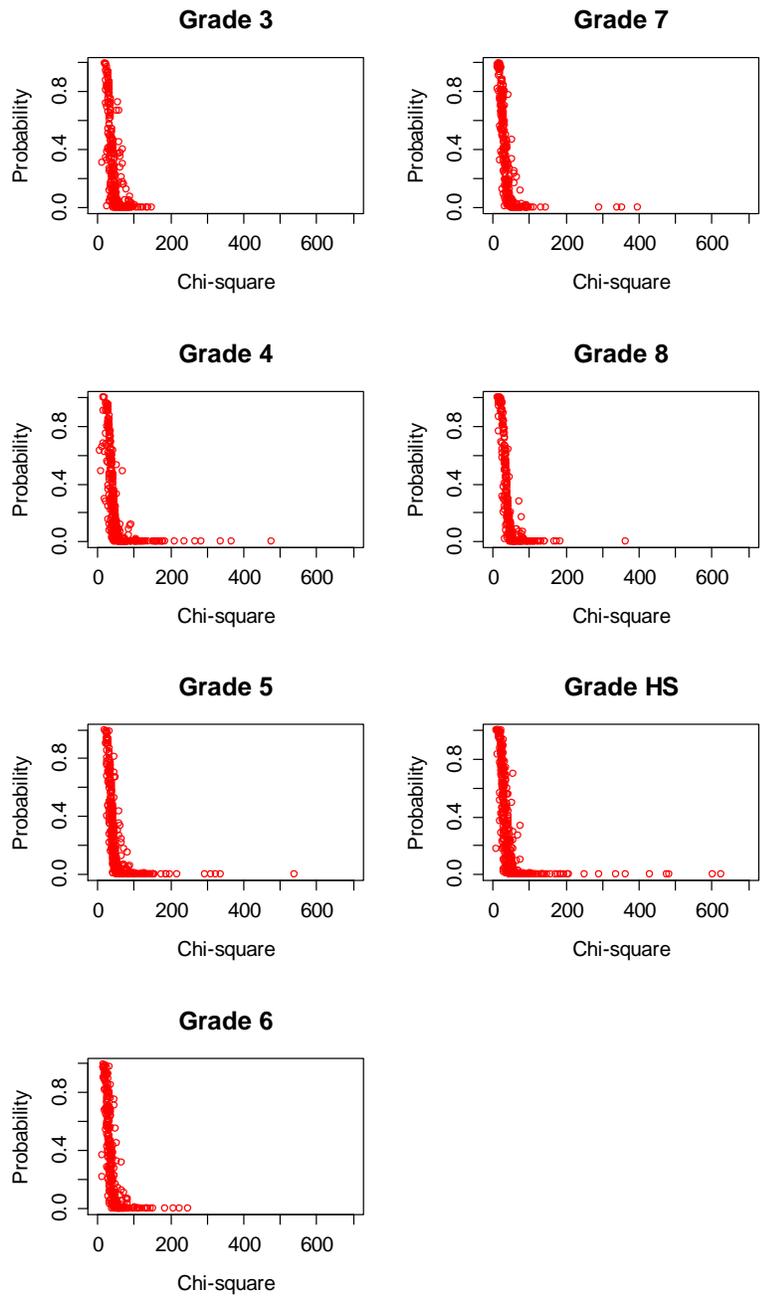


Figure 8. Mathematics Item Fit Chi-Square Plots (Vertical Scaling)

Table 10. Summary of Likelihood Ratio χ^2 Test Statistics by Grade and Content Area.

Grade	No. of Students per Item (Range)	Mean	SD	Min.	Max.	Prob.<.05	Prob.<.01
ELA/literacy							
3	530 - 9,804	50	27	15	199	68	41
4	550 - 22,291	51	29	7	244	102	57
5	759 - 10,853	50	30	3	270	78	44
6	527 - 13,746	56	37	9	335	87	53
7	509 - 20,748	57	42	9	367	83	54
8	508 - 12,981	57	37	1	297	104	57
HS	526 - 16,646	58	47	9	497	156	98
Mathematics							
3	693 - 5,952	49	21	14	150	106	61
4	519 - 13,845	60	46	7	475	155	107
5	502 - 19,614	62	47	19	538	166	119
6	509 - 7,722	47	31	11	247	98	71
7	502 - 10,188	44	40	12	395	97	49
8	536 - 13,005	52	33	12	364	119	79
HS	501 - 14,521	56	67	9	626	122	78

Table 11. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4.

No. of Items = 120	<i>a</i> -parameter		<i>b</i> -parameter	
	3	4	3	4
Mean	0.62	0.63	0.87	0.35
SD	0.21	0.20	1.14	1.12
Min	0.12	0.17	-1.09	-1.54
10%	0.36	0.39	-0.55	-1.06
25%	0.47	0.51	-0.11	-0.55
Median	0.63	0.61	0.86	0.32
75%	0.73	0.74	1.58	1.07
90%	0.91	0.87	2.29	1.87
Max	1.19	1.32	3.76	3.22

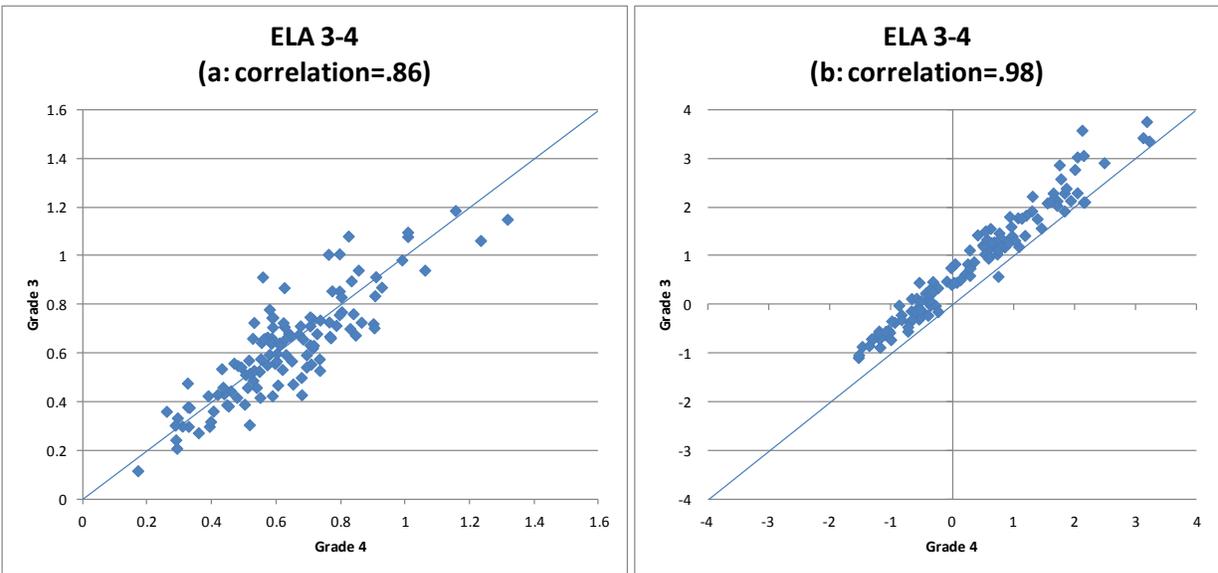


Figure 9. Comparison of ELA/literacy *a* and *b*-parameter estimates for Linking Grade 3 to 4

Table 12. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5.

No. of Items = 133	<i>a</i> -parameter		<i>b</i> -parameter	
	4	5	4	5
Mean	0.60	0.60	0.79	0.37
SD	0.21	0.19	1.15	1.14
Min	0.22	0.18	-2.33	-3.10
10%	0.31	0.36	-0.63	-0.89
25%	0.43	0.45	0.08	-0.31
Median	0.59	0.62	0.63	0.21
75%	0.76	0.73	1.50	1.07
90%	0.84	0.82	2.52	2.18
Max	1.13	1.17	3.61	3.28

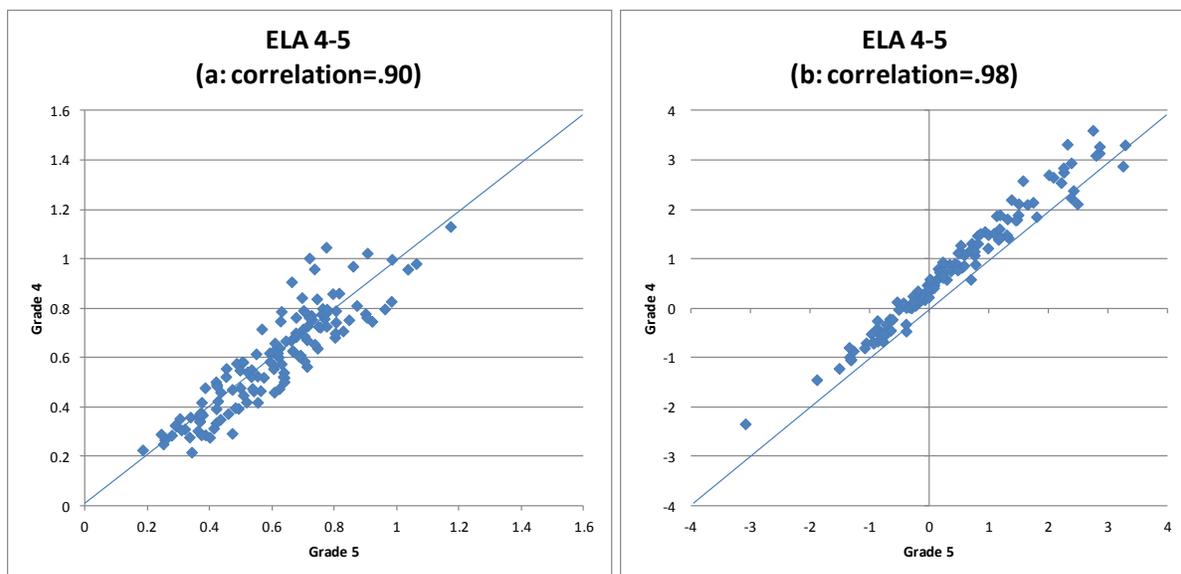

 Figure 10. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 4 to 5

Table 13. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6.

No. of Items = 131	<i>a</i> -parameter		<i>b</i> -parameter	
	5	6	5	6
Mean	0.61	0.60	0.75	0.53
SD	0.21	0.22	1.17	1.16
Min	0.14	0.17	-2.07	-2.00
10%	0.35	0.36	-0.82	-0.98
25%	0.46	0.46	-0.04	-0.24
Median	0.60	0.57	0.66	0.49
75%	0.74	0.72	1.69	1.52
90%	0.87	0.87	2.29	2.06
Max	1.22	1.32	3.25	3.00

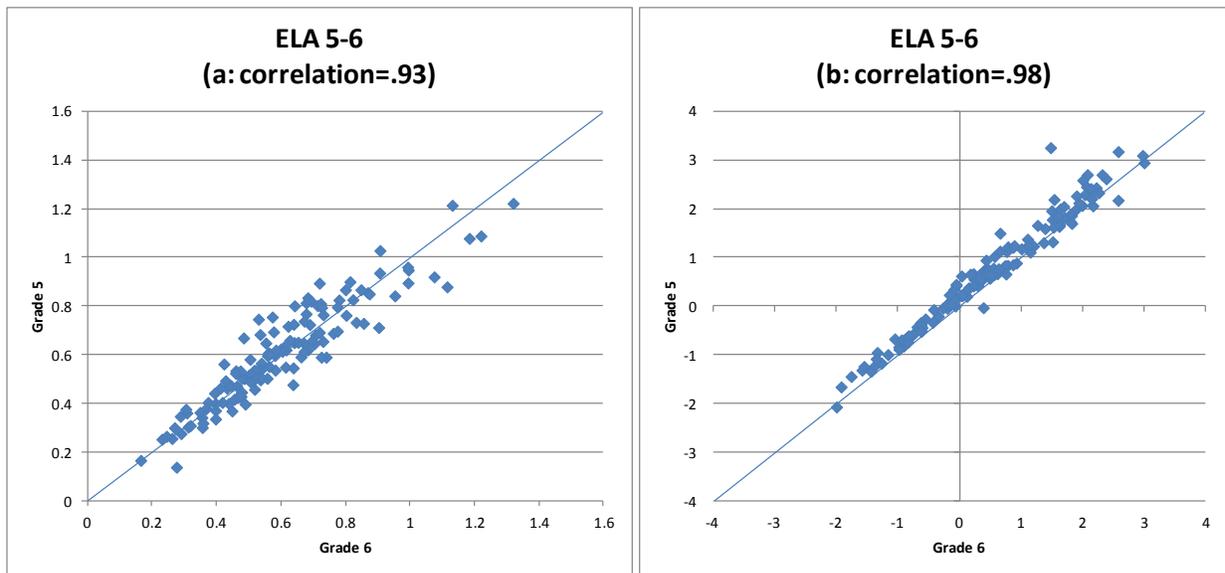


Figure 11. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 5 to 6

Table 14. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6.

No. of Items = 107	<i>a</i> -parameter		<i>b</i> -parameter	
	6	7	6	7
Mean	0.59	0.60	1.06	0.84
SD	0.20	0.19	1.22	1.13
Min	0.18	0.21	-1.88	-2.02
10%	0.34	0.32	-0.41	-0.60
25%	0.44	0.46	0.19	0.07
Median	0.57	0.59	1.15	0.96
75%	0.73	0.73	1.86	1.58
90%	0.85	0.83	2.60	2.18
Max	1.01	1.09	4.48	3.50

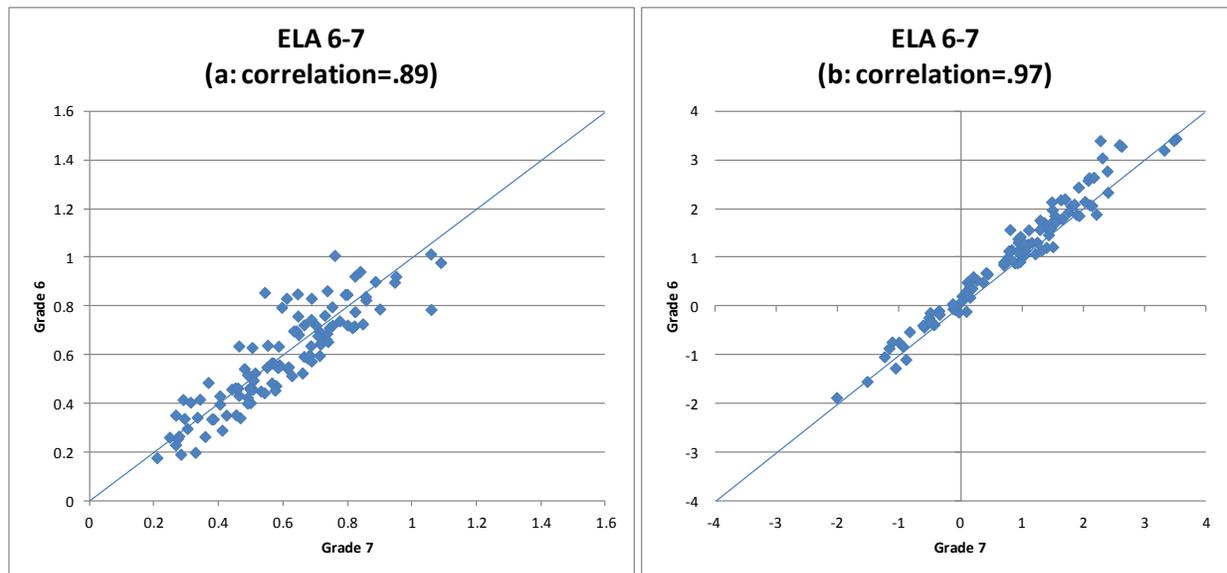


Figure 12. Comparison of ELA/literacy *a*- and *b*-parameter estimates for linking Grade 7 to 6

Table 15. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7.

No. of Items = 123	<i>a</i> -parameter		<i>b</i> -parameter	
	7	8	7	8
Mean	0.59	0.59	1.01	0.77
SD	0.21	0.20	1.23	1.23
Min	0.19	0.19	-2.20	-2.42
10%	0.32	0.34	-0.34	-0.65
25%	0.44	0.44	0.09	-0.16
Median	0.56	0.56	0.98	0.79
75%	0.72	0.70	1.83	1.52
90%	0.85	0.88	2.53	2.12
Max	1.22	1.09	4.61	4.16

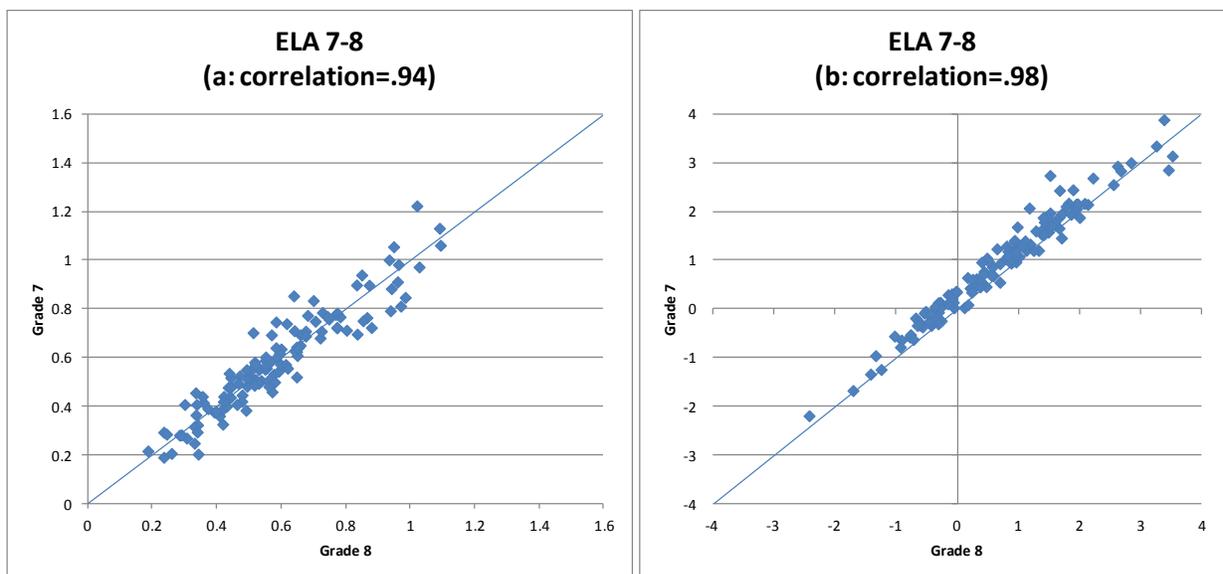

 Figure 13. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 8 to 7

Table 16. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: High School to Grade 8.

No. of Items = 107	<i>a</i> -parameter		<i>b</i> -parameter	
	8	HS	8	HS
Mean	0.57	0.60	1.16	0.93
SD	0.25	0.26	1.43	1.27
Min	0.12	0.13	-2.01	-2.02
10%	0.28	0.30	-0.58	-0.69
25%	0.38	0.40	0.05	0.03
Median	0.57	0.59	1.05	0.86
75%	0.70	0.77	2.09	1.63
90%	0.88	0.94	3.18	2.57
Max	1.43	1.35	4.23	4.20

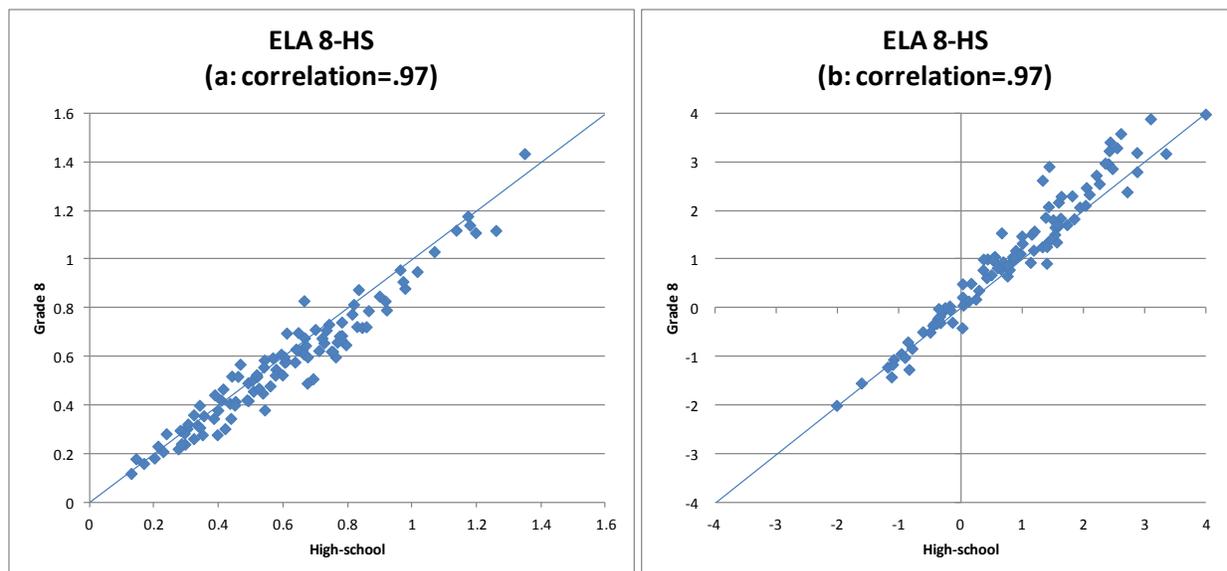

 Figure 14. Comparison of ELA/literacy *a* and *b*-parameter estimates for Linking High School to Grade 8

Table 17. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4.

No. of Items = 104	<i>a</i> -parameter		<i>b</i> -parameter	
	3	4	3	4
Mean	0.77	0.81	0.61	-0.07
SD	0.25	0.23	1.30	1.17
Min	0.15	0.26	-2.17	-2.88
10%	0.44	0.52	-0.97	-1.49
25%	0.58	0.63	-0.42	-0.92
Median	0.77	0.81	0.65	-0.07
75%	0.95	1.00	1.39	0.67
90%	1.09	1.09	2.23	1.50
Max	1.38	1.29	4.83	3.47

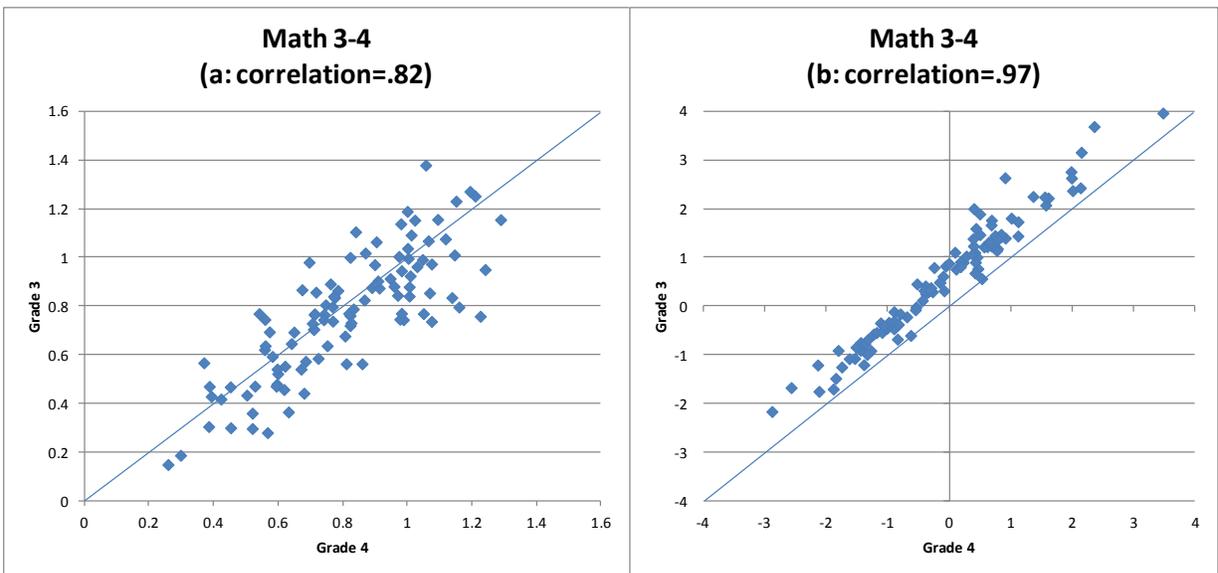

 Figure 15. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 3 to 4

Table 18. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5.

No. of Items = 95	<i>a</i> -parameter		<i>b</i> -parameter	
	4	5	4	5
Mean	0.77	0.81	0.87	0.45
SD	0.26	0.25	0.98	0.95
Min	0.33	0.35	-1.73	-1.81
10%	0.42	0.49	-0.52	-0.87
25%	0.58	0.65	0.17	-0.32
Median	0.75	0.79	0.89	0.43
75%	0.94	0.94	1.57	1.24
90%	1.11	1.16	2.17	1.65
Max	1.46	1.49	2.88	2.80

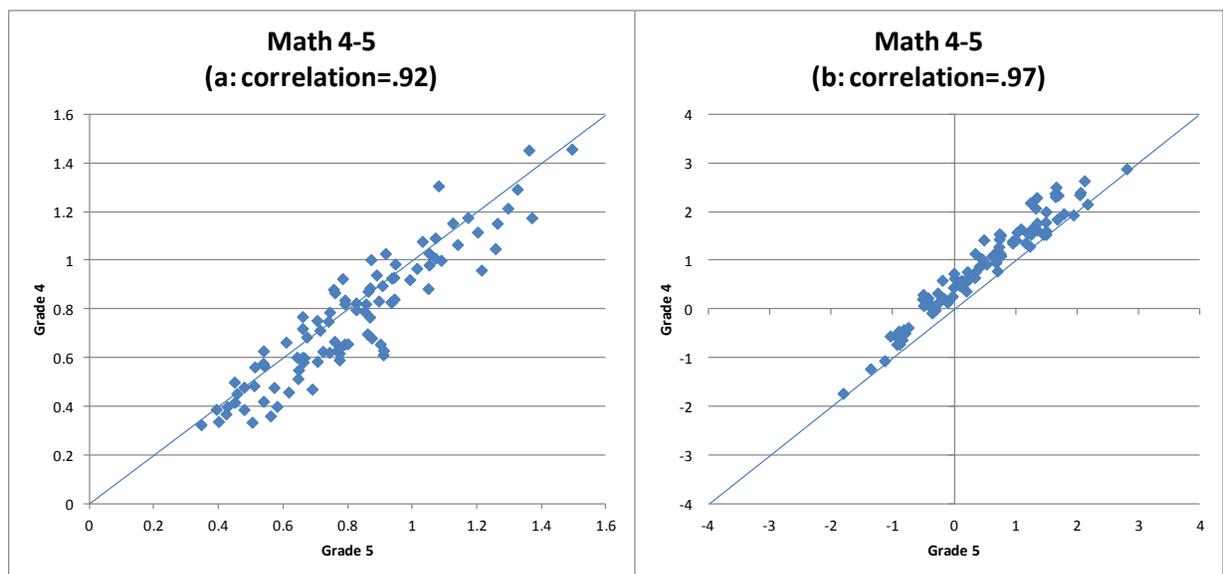

 Figure 16. Comparison of Mathematics *a* and *b*-parameter estimates for Linking Grade 4 to 5

Table 19. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6.

No. of Items = 102	<i>a</i> -parameter		<i>b</i> -parameter	
	5	6	5	6
Mean	0.69	0.69	0.84	0.58
SD	0.28	0.25	1.00	1.05
Min	0.26	0.23	-1.78	-1.95
10%	0.40	0.40	-0.45	-0.76
25%	0.49	0.53	0.19	-0.21
Median	0.62	0.64	0.89	0.58
75%	0.85	0.85	1.40	1.37
90%	1.06	1.01	2.17	1.93
Max	1.61	1.47	3.14	3.82

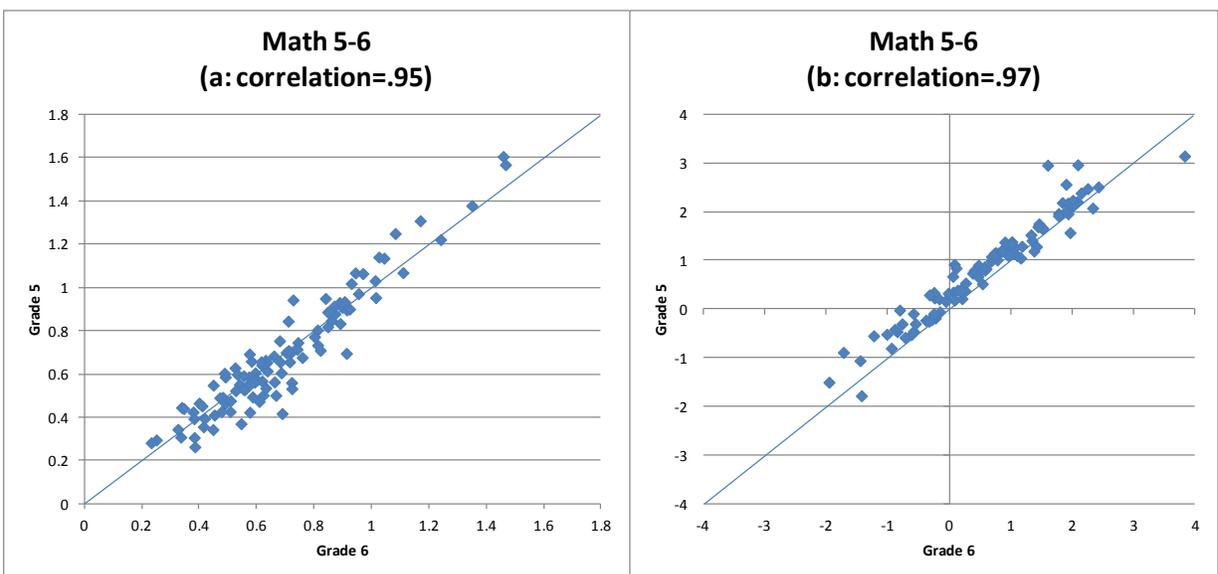

 Figure 17. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 5 to 6

Table 20. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6.

No. of Items = 71	<i>a</i> -parameter		<i>b</i> -parameter	
	6	7	6	7
Mean	0.76	0.85	1.04	0.77
SD	0.30	0.31	1.11	1.00
Min	0.13	0.20	-1.74	-1.70
10%	0.36	0.48	-0.22	-0.47
25%	0.56	0.65	0.40	0.22
Median	0.76	0.86	1.01	0.74
75%	0.97	1.12	1.87	1.52
90%	1.14	1.22	2.37	1.99
Max	1.59	1.37	3.28	2.62

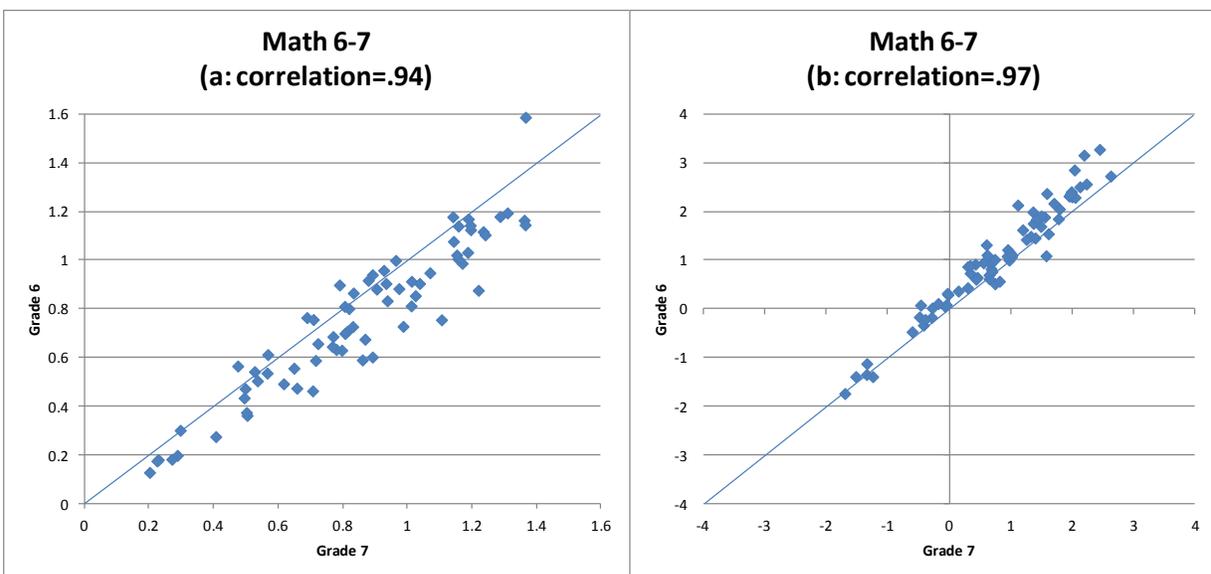

 Figure 18. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 7 to 6

Table 21. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7.

No. of Items = 73	<i>a</i> -parameter		<i>b</i> -parameter	
	7	8	7	8
Mean	0.84	0.86	1.18	0.97
SD	0.34	0.34	1.20	1.20
Min	0.22	0.22	-1.66	-1.80
10%	0.34	0.38	-0.17	-0.35
25%	0.59	0.63	0.36	0.10
Median	0.88	0.84	1.06	1.10
75%	1.10	1.14	1.87	1.65
90%	1.23	1.33	2.80	2.41
Max	1.80	1.60	4.33	4.13

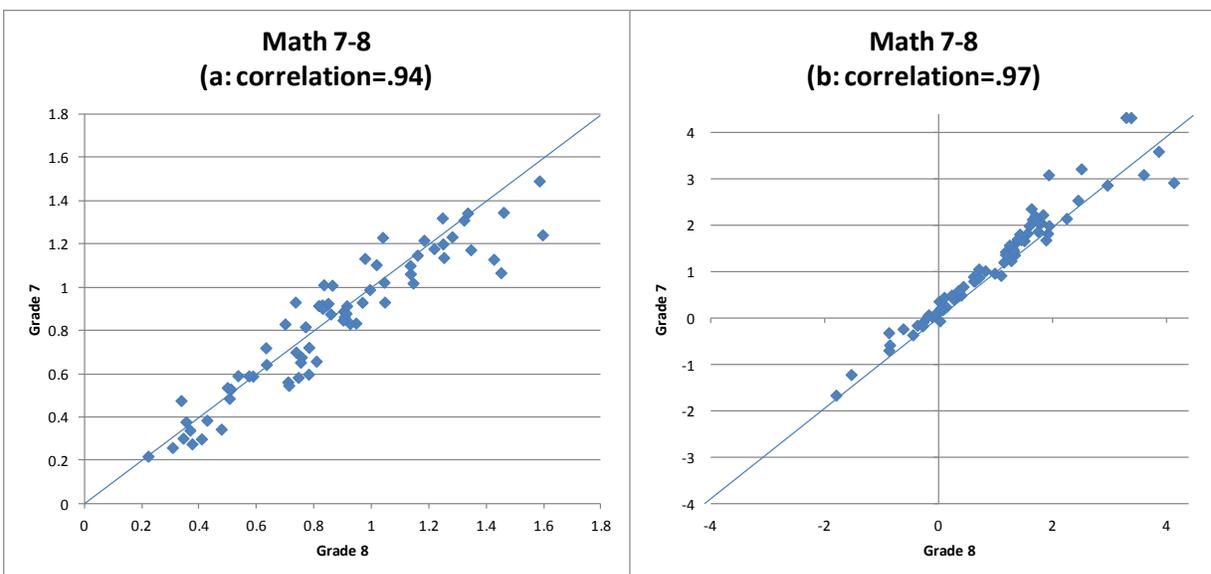

 Figure 19. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 8 to 7

Table 22. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: High School to Grade 8.

No. of Items = 81	<i>a</i> -parameter		<i>b</i> -parameter	
	8	HS	8	HS
Mean	0.67	0.76	1.33	0.87
SD	0.31	0.32	1.34	1.12
Min	0.15	0.26	-1.89	-1.47
10%	0.30	0.37	-0.10	-0.43
25%	0.40	0.50	0.46	0.21
Median	0.62	0.76	1.45	0.92
75%	0.93	1.00	2.09	1.50
90%	1.09	1.16	3.33	2.52
Max	1.32	1.48	4.16	3.39

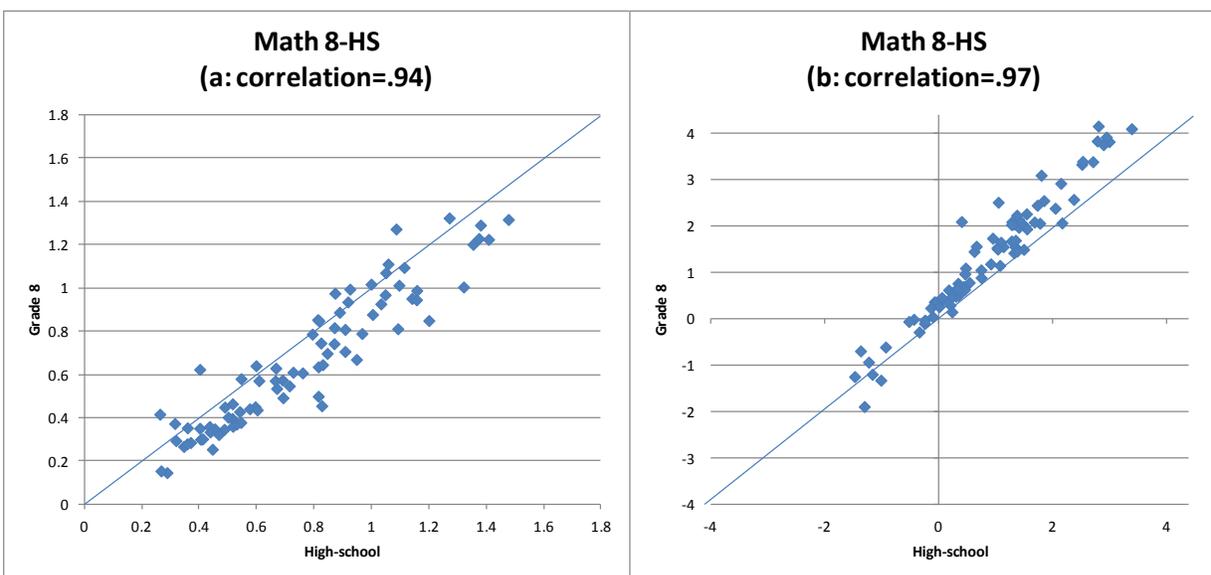

 Figure 20. Comparison of Mathematics *a* and *b*-parameter estimates for Linking High School to Grade 8

Table 23. Vertical Linking Transformation Constants from the Stocking-Lord Procedure.

Grade Pairs	Slope A	Intercept B
ELA/literacy		
3 to 4	0.944421	-1.188941
4 to 5	0.973260	-0.683668
5 to 6	1.002164	-0.256198
6 (Base Grade)		
7 to 6	1.027782	0.172946
8 to 7	1.033673	0.437905
HS to 8	1.105322	0.583021
Mathematics		
3 to 4	0.872487	-1.240565
4 to 5	0.938657	-0.666856
5 to 6	1.004384	-0.279283
6 (Base Grade)		
7 to 6	1.103163	0.147206
8 to 7	1.137342	0.340534
HS to 8	1.311837	0.630426

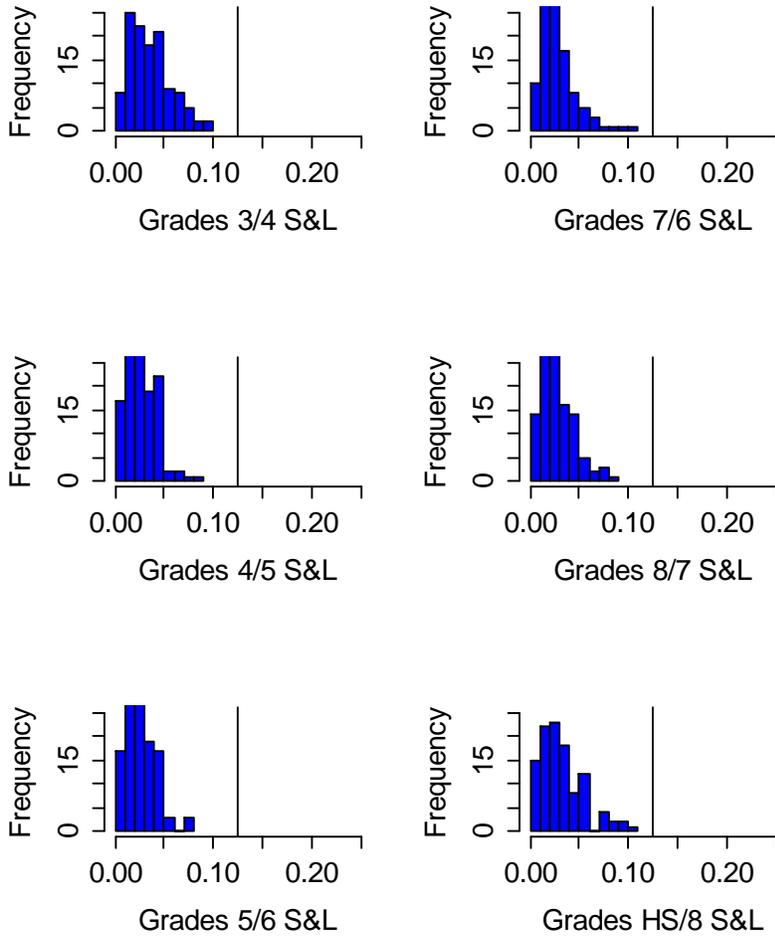


Figure 21. Distribution of WRMSD for ELA/literacy (Vertical Linking Items)

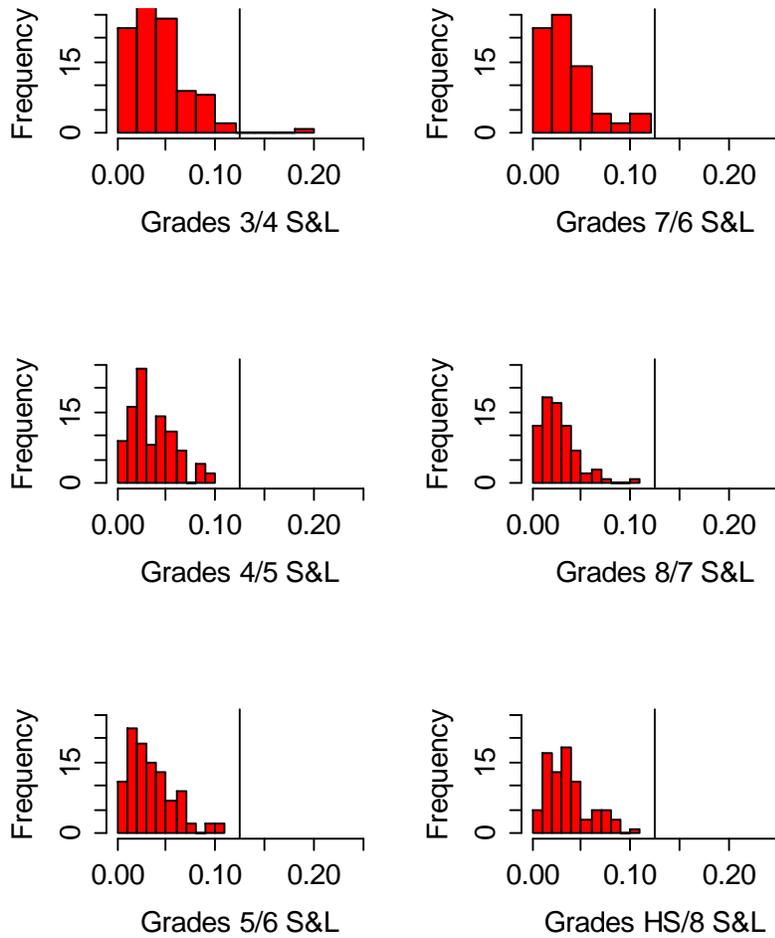


Figure 22. Distribution of WRMSD for Mathematics (Vertical Linking Items)

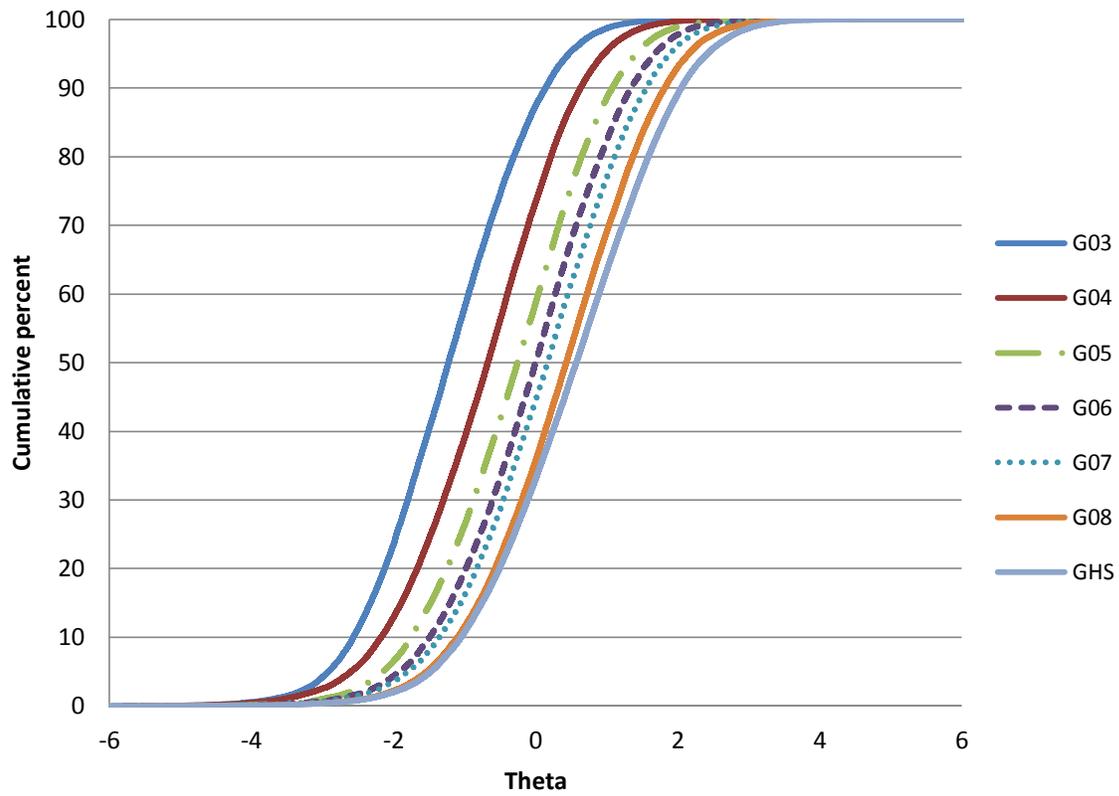


Figure 23. ELA/literacy Cumulative Distributions of Student Ability across Grades

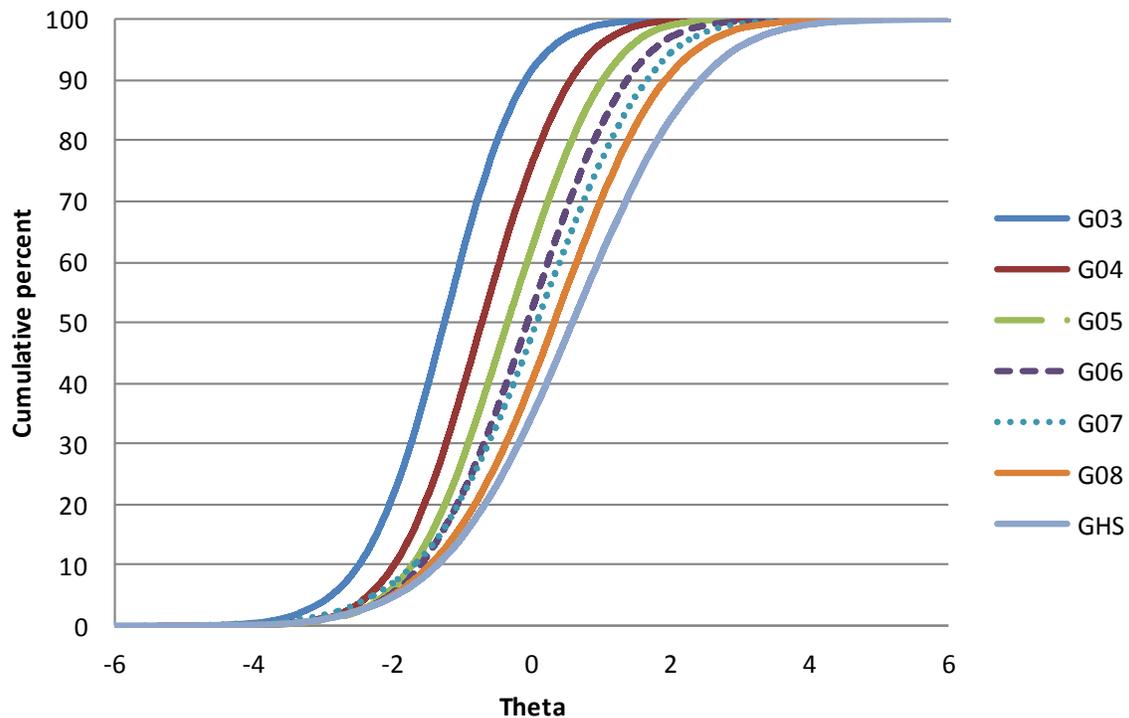


Figure 24. Mathematics Cumulative Distributions of Student Ability across Grades

Figures 25 and 26 present plots of the univariate theta (i.e., student ability) distributions for ELA/literacy and mathematics. Figures 27 and 28 present the ability distributions using boxplots for ELA/literacy and mathematics. The boxplots show that both the means and standard deviations tend to increase with grade level. No constraints were placed on the minimum and maximum thetas in the box plots (refer to the section on establishing the minimum and maximum thetas). The properties of the vertical scale are consistent with the comments of Kolen (2011) that an acceptable vertical scale should display increasing mean scores from grade-to-grade, the amount of growth should be decelerating, and the within-grade variability (SD) should be increasing from grade to grade.

Table 24. Summary of Vertically Scaled Student Ability Estimates and Effect Size.

Grade	<i>N</i>	Mean	<i>SD</i>	10%	25%	Median	75%	90%	Effect Size
ELA/literacy									
3	23,223	-1.227	1.051	-2.568	-1.961	-1.229	-0.487	0.138	
4	35,689	-0.737	1.106	-2.176	-1.476	-0.674	0.054	0.631	0.45
5	31,594	-0.305	1.102	-1.752	-1.052	-0.254	0.479	1.075	0.39
6	31,535	-0.048	1.107	-1.491	-0.785	-0.002	0.731	1.342	0.23
7	30,913	0.119	1.139	-1.357	-0.633	0.166	0.933	1.555	0.15
8	35,913	0.385	1.142	-1.099	-0.377	0.432	1.197	1.816	0.23
HS	50,657	0.527	1.211	-1.050	-0.301	0.573	1.400	2.050	0.12
Mathematics									
3	24,799	-1.265	0.960	-2.480	-1.872	-1.249	-0.634	-0.079	
4	38,925	-0.700	0.997	-1.972	-1.367	-0.705	-0.020	0.581	0.58
5	42,380	-0.330	1.073	-1.699	-1.052	-0.327	0.421	1.048	0.36
6	29,946	-0.083	1.183	-1.586	-0.858	-0.049	0.726	1.391	0.22
7	28,271	0.030	1.336	-1.691	-0.821	0.083	0.945	1.681	0.09
8	34,880	0.276	1.333	-1.469	-0.582	0.324	1.183	1.926	0.18
HS	47,608	0.571	1.494	-1.351	-0.408	0.598	1.578	2.448	0.21

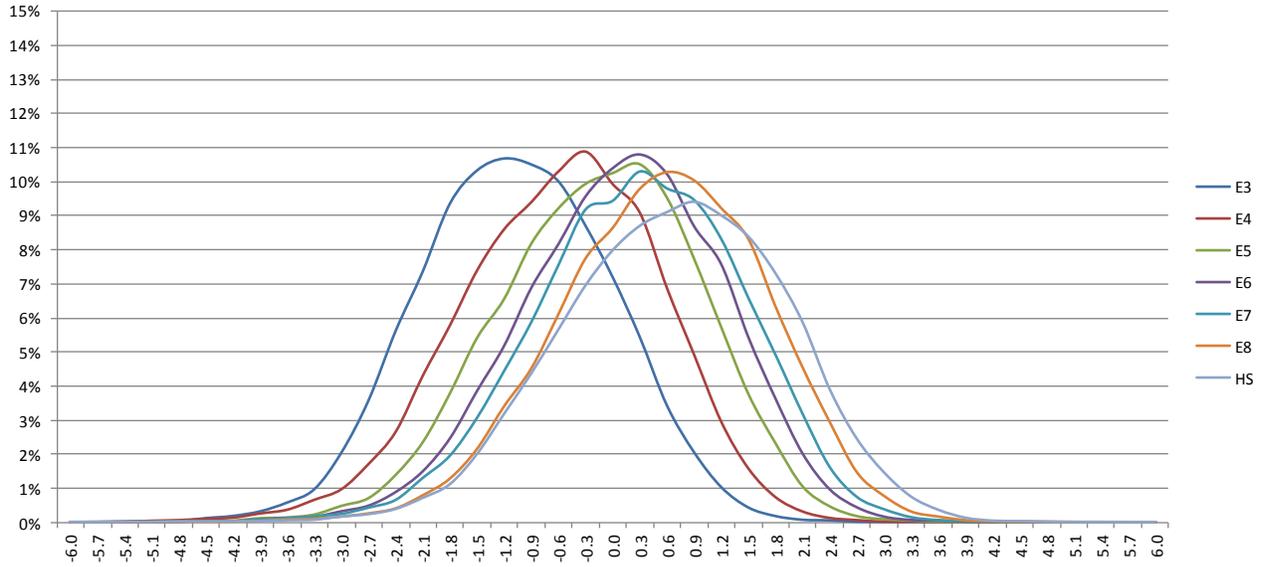


Figure 25. ELA/literacy Student Ability Distributions Across Grades 3 to High School

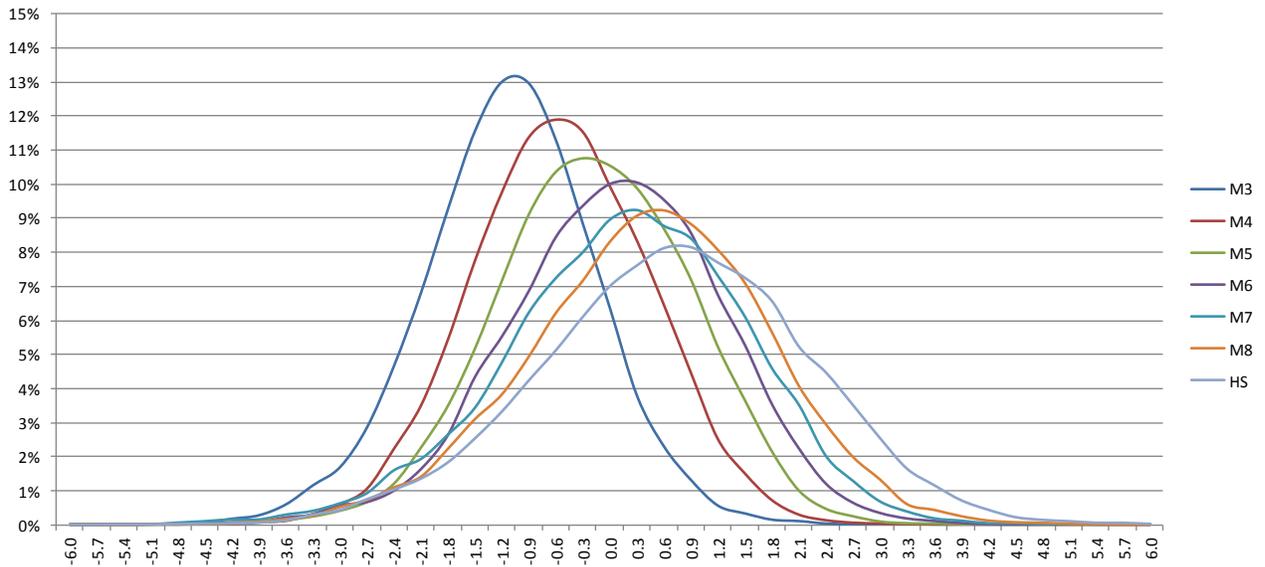


Figure 26. Mathematics Student Ability Distributions Across Grades 3 to High School

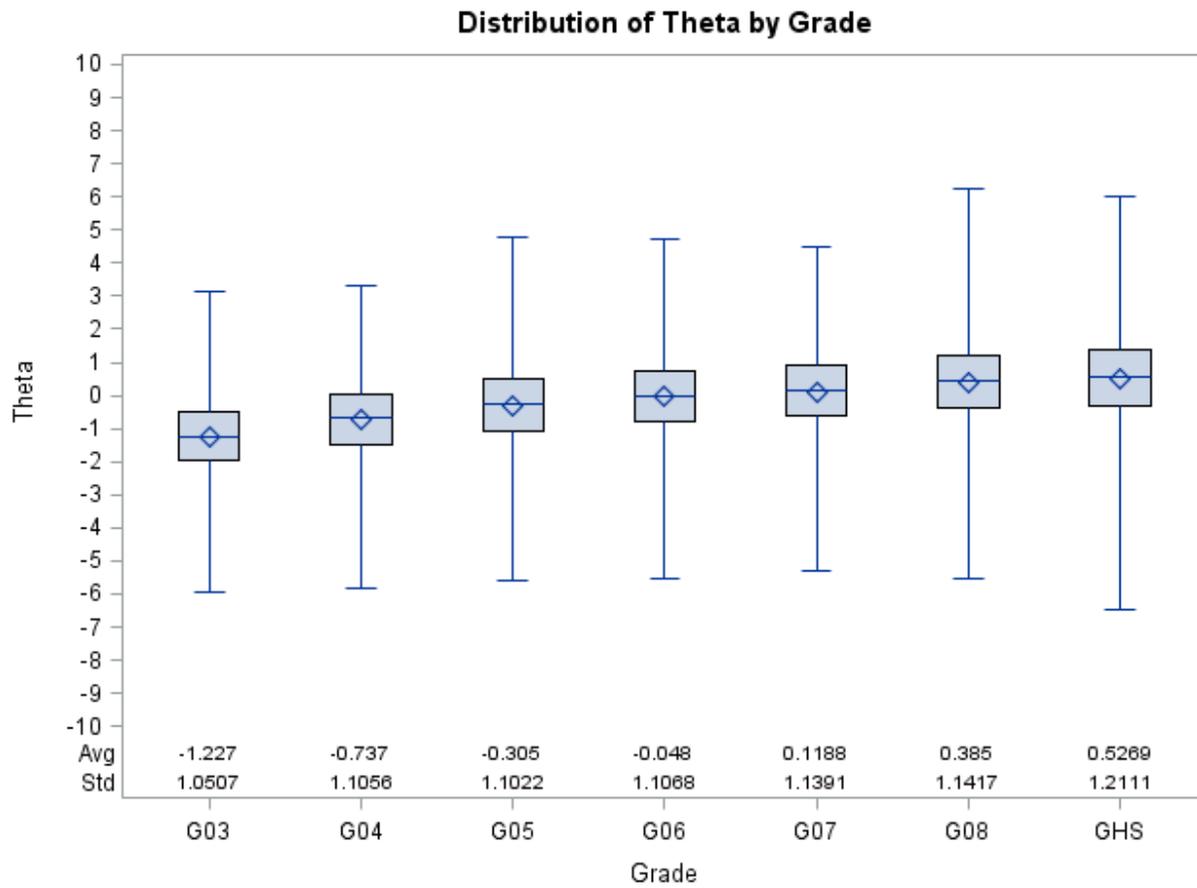


Figure 27. Boxplots of Theta Estimates across Grade Level for ELA/literacy

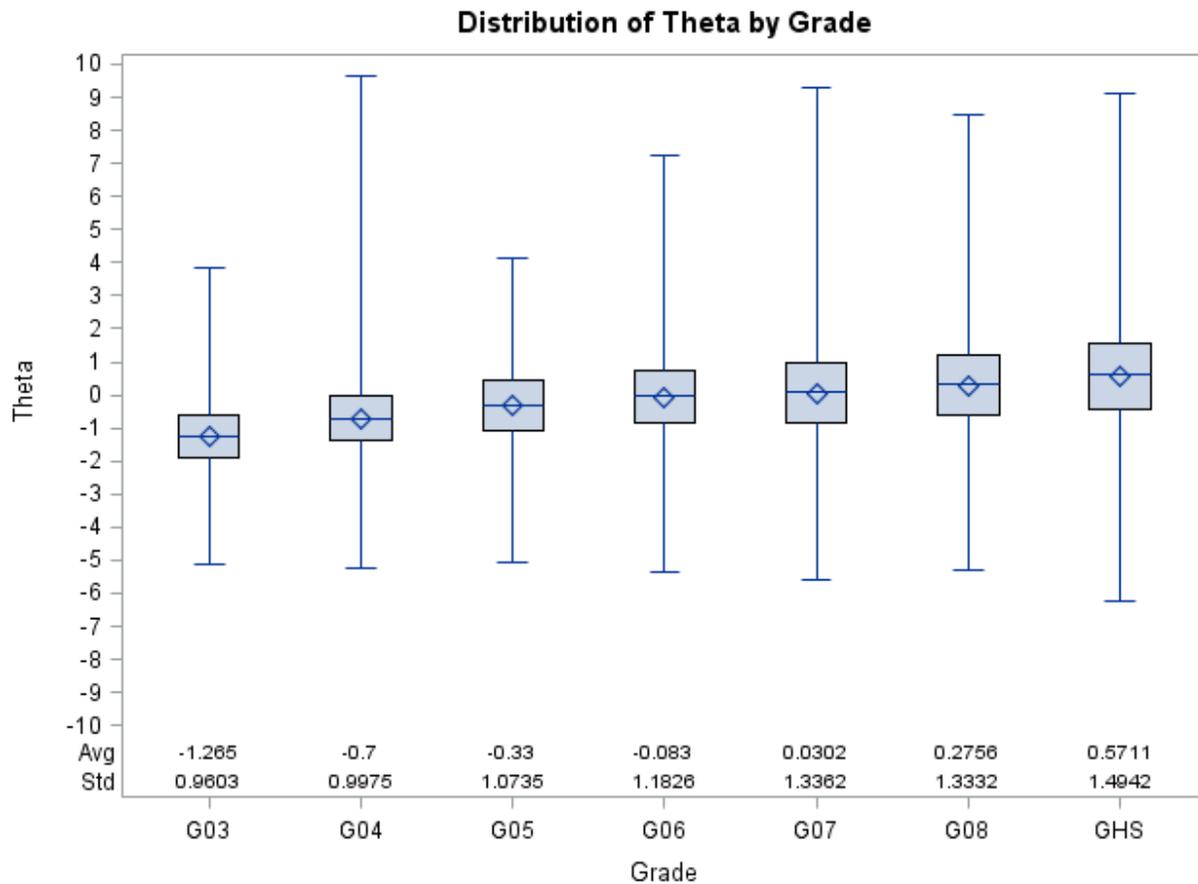


Figure 28. Boxplots of Theta Estimates Across Grade Level for Mathematics

Figures 29 to 34 display IRT information functions by score level and the total combined across all score levels (i.e., All) for ELA/literacy and mathematics. These displays reflect all items, both on-grade and off-grade vertical linking items, administered in a given grade level for the vertical scaling. Figures 31 and 34 show total information across grades for ELA/literacy and mathematics.

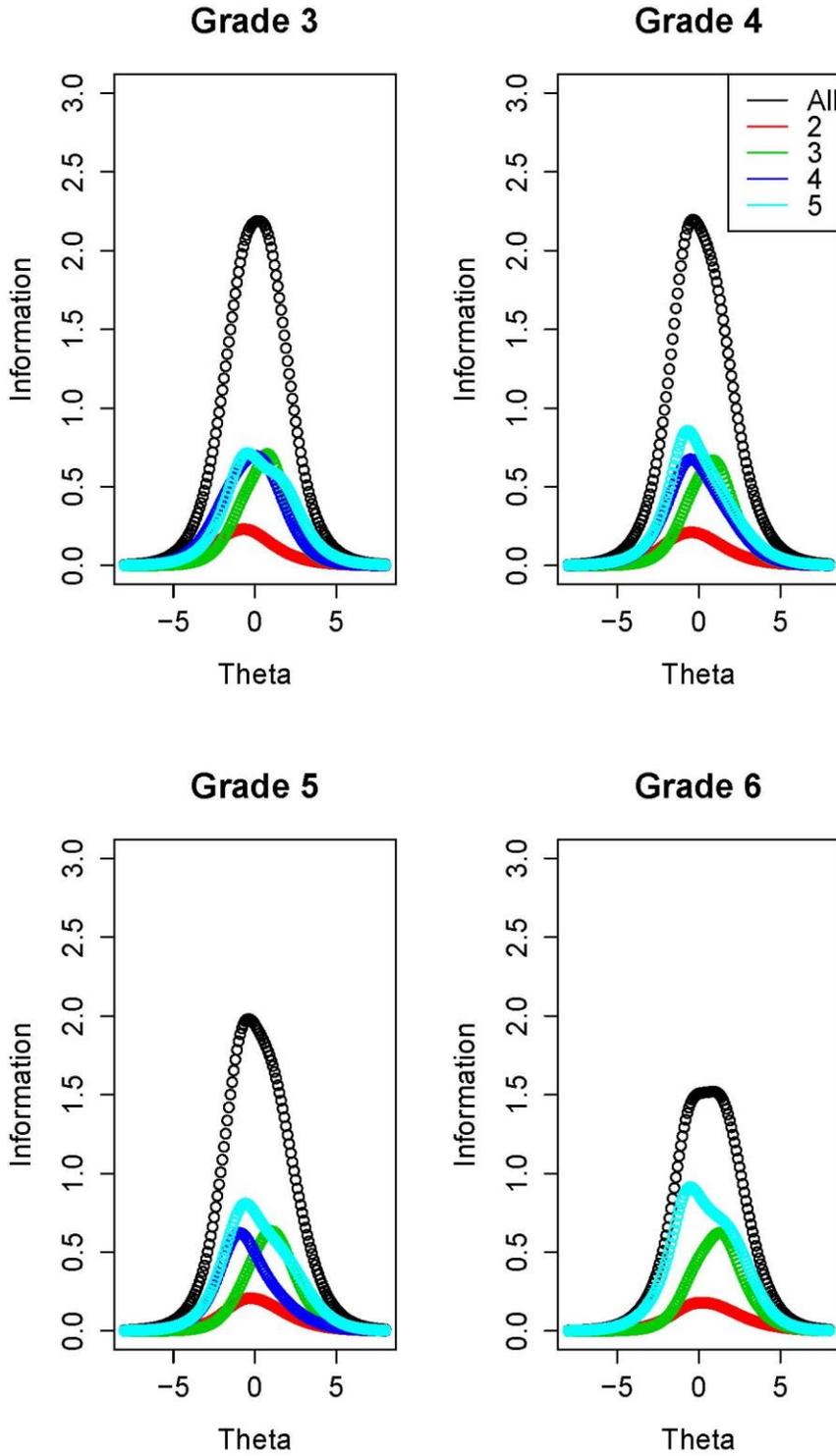


Figure 29. ELA/literacy Test Information by Score Level and Combined for Grades 3 to 6

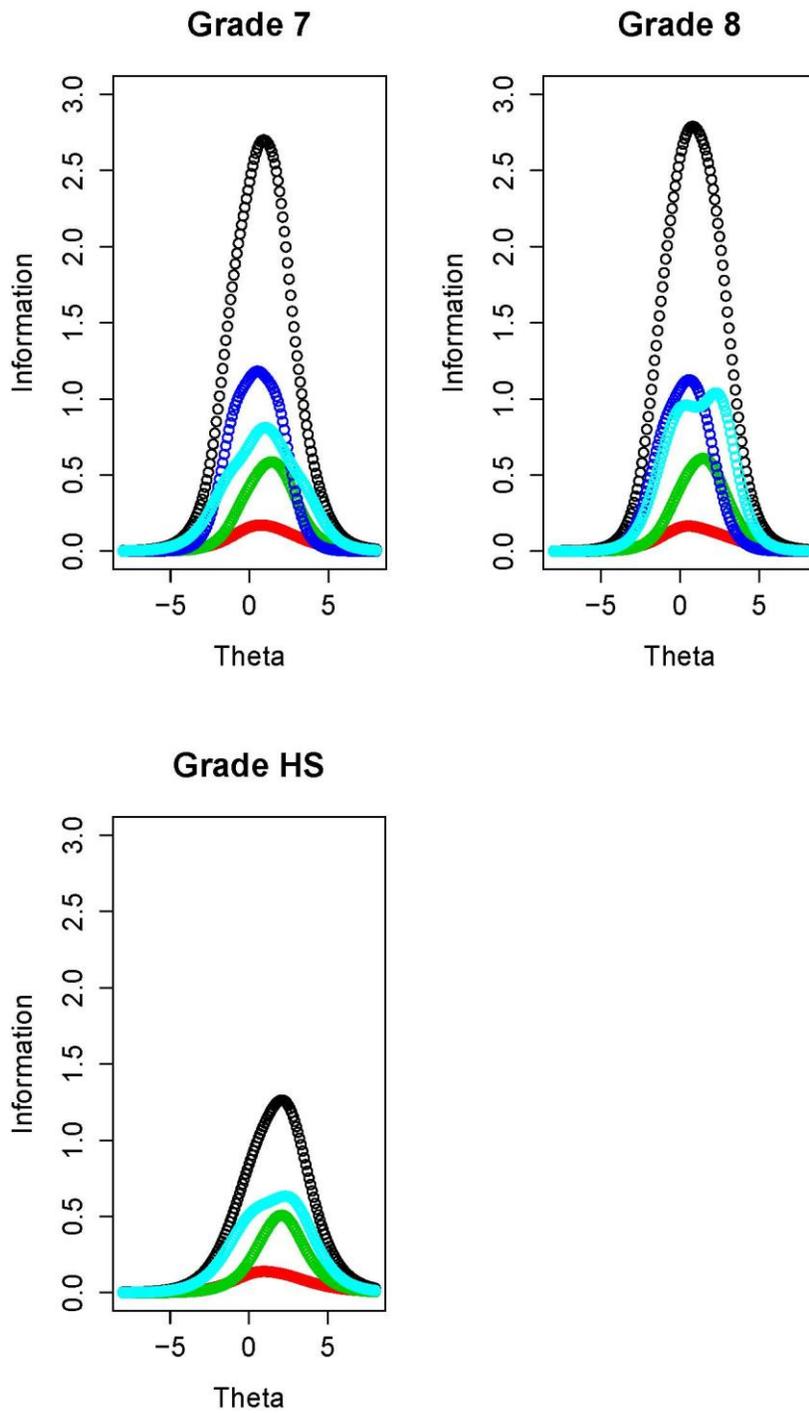


Figure 30. ELA/literacy Test Information by Score Level and Combined for Grades 7 to High School

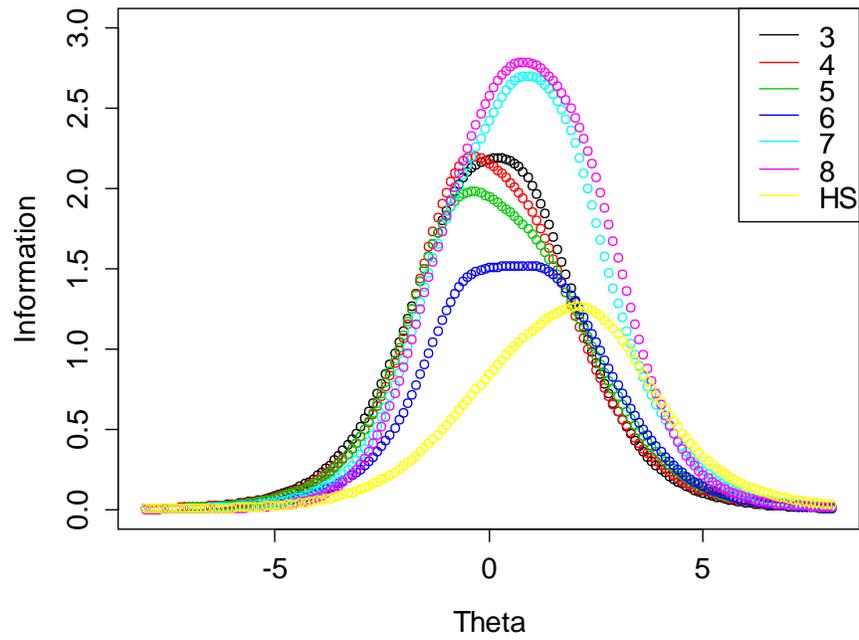


Figure 31. ELA/literacy Total Test Information for Grades 3 to High School

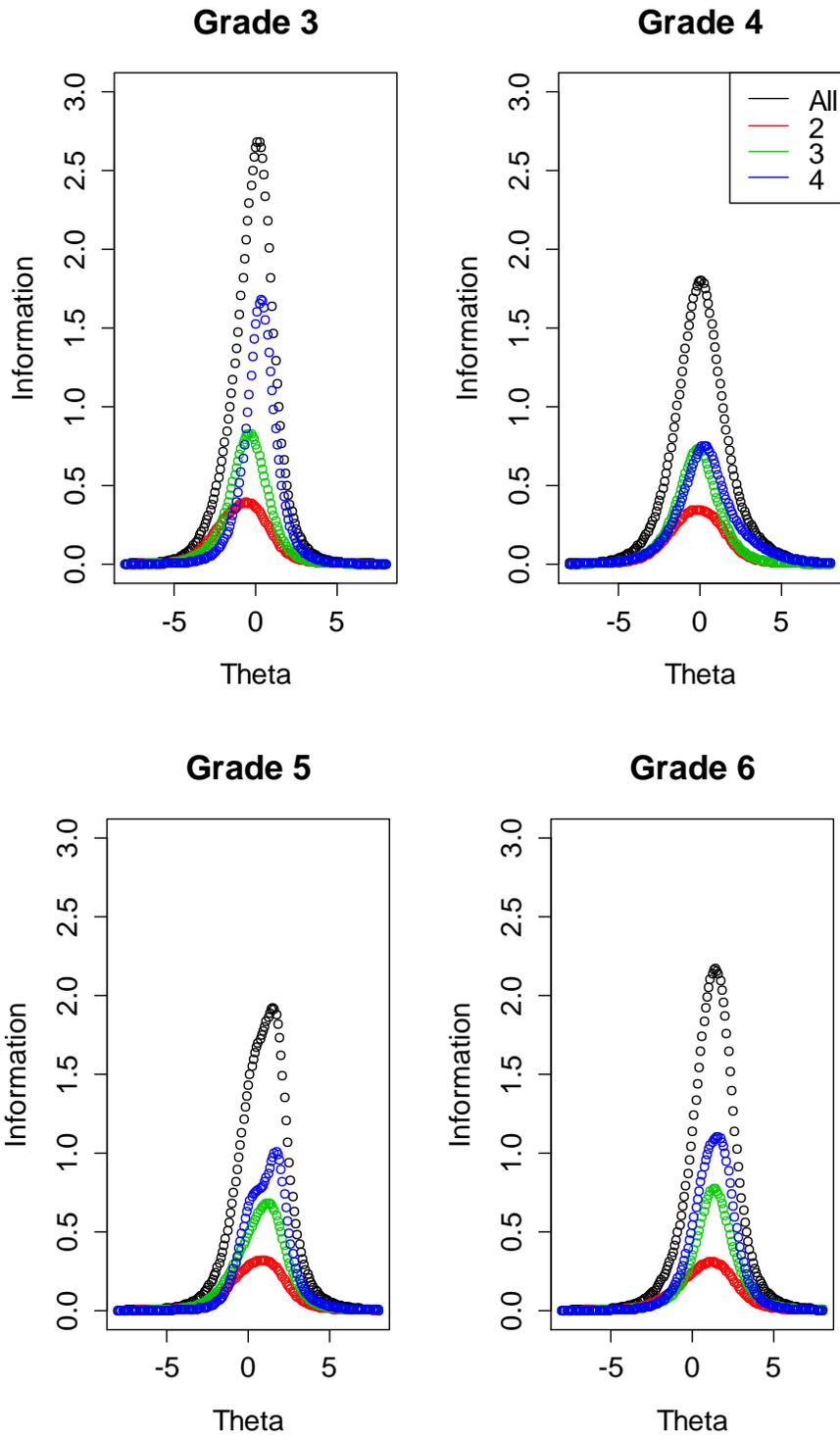


Figure 32. Mathematics Test Information and Score Level and Combined for Grades 3 to 6

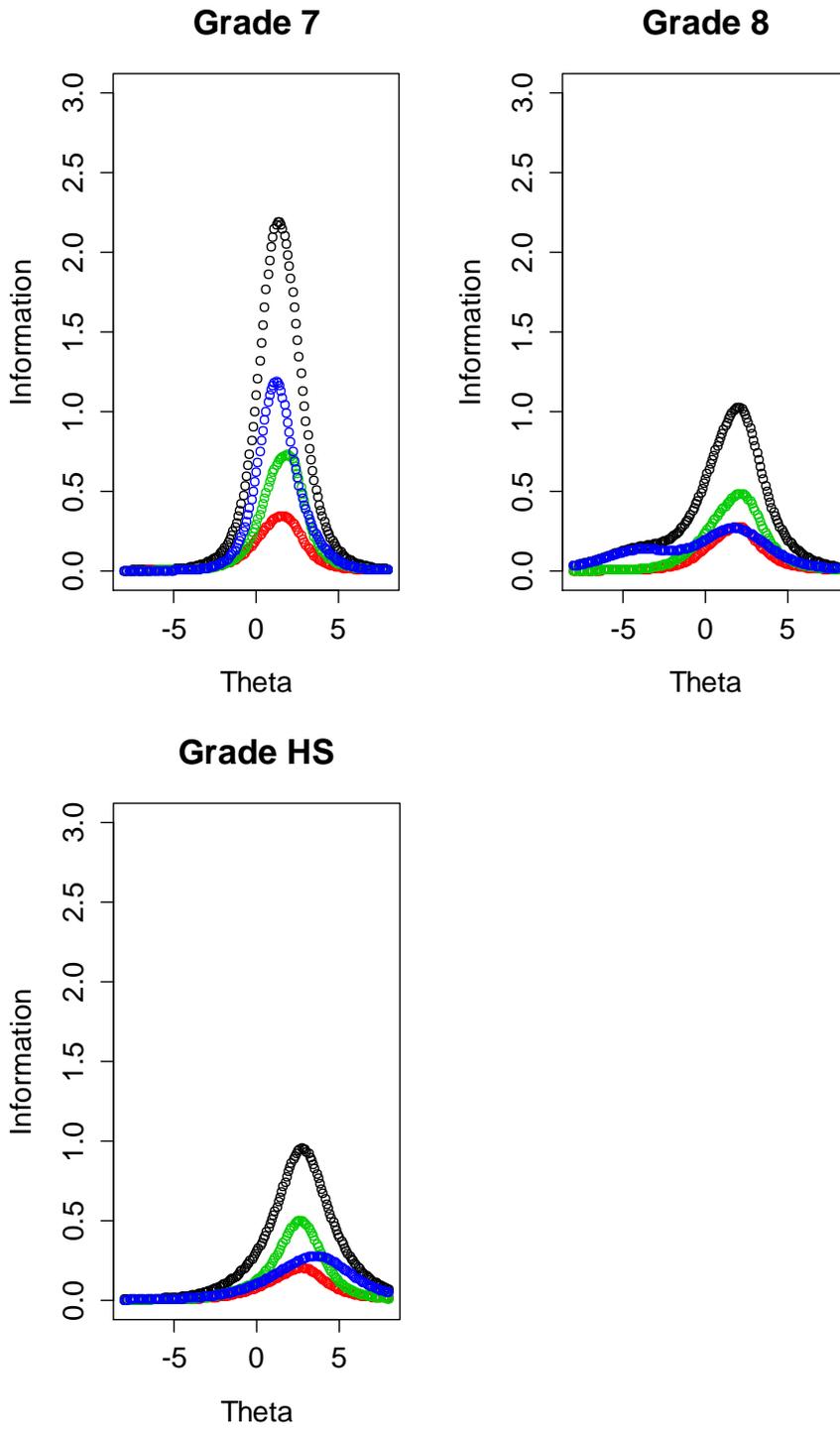


Figure 33. Mathematics Test Information and Score Level and Combined for Grades 7 to High School

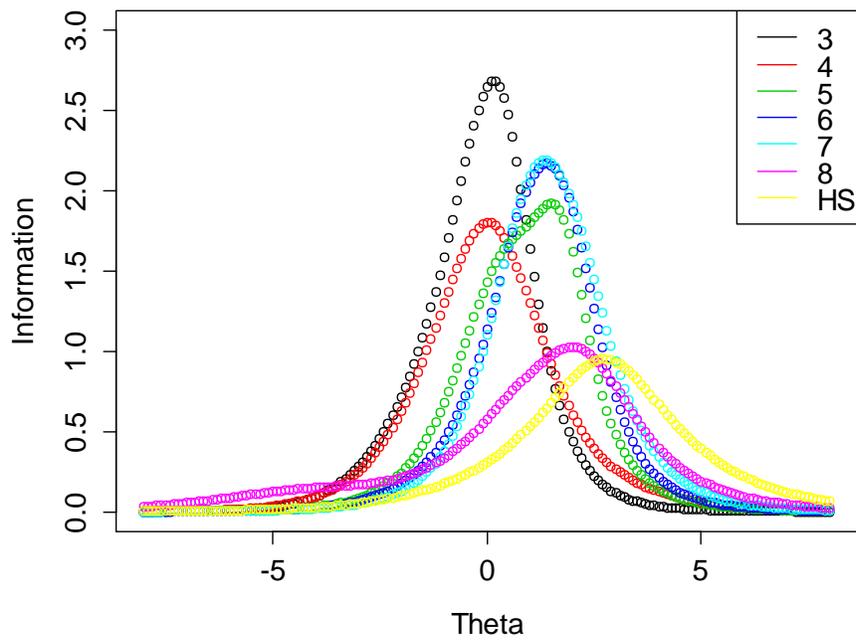


Figure 34. Mathematics Total Test Information for Grades 3 to High School

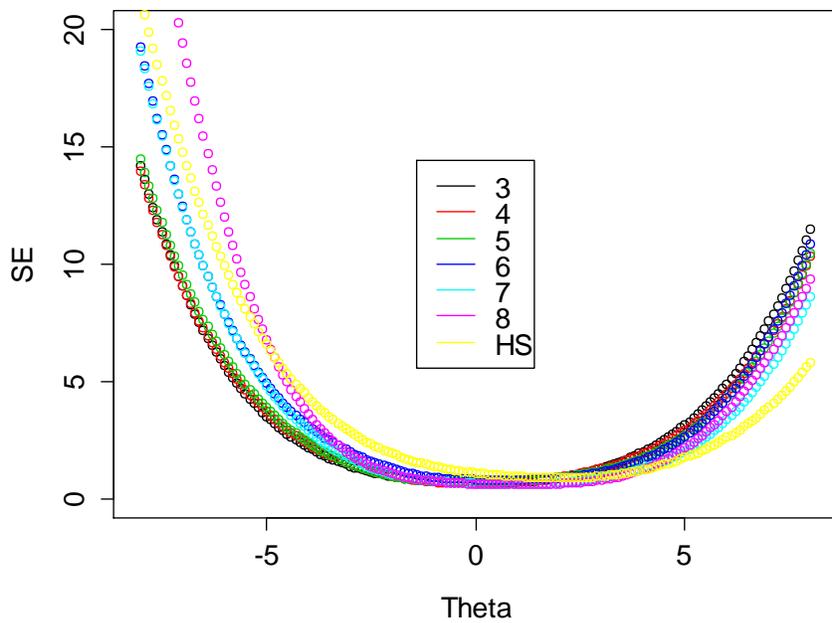


Figure 35. IRT Standard Error Plots for ELA/literacy Grades 3 to High School (HS)

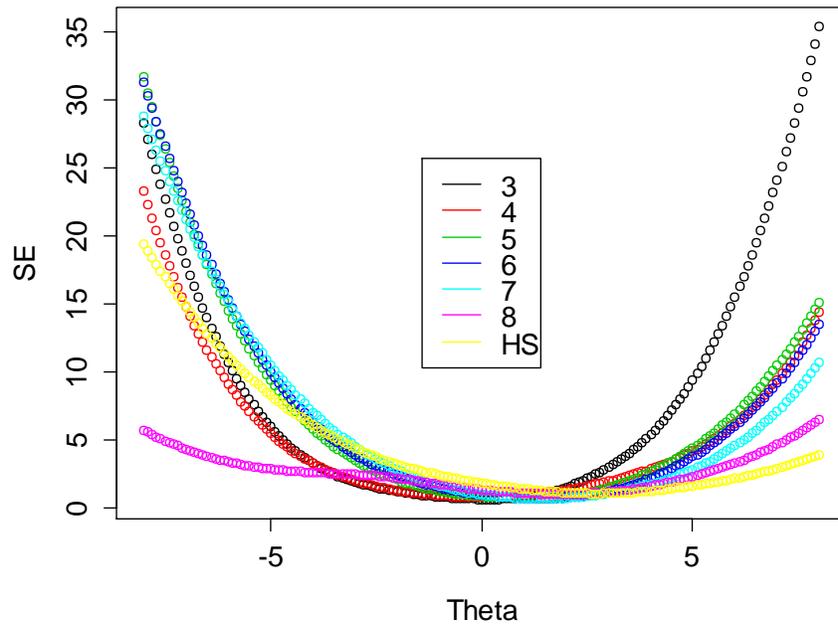


Figure 36. IRT Standard Error Plots for Mathematics Grades 3 to High School (HS)

IRT conditional standard errors were calculated as one over the inverse of information for a given level of theta. Plots of the conditional standard errors of measurement (*CSEM*) for ELA/literacy and mathematics are displayed in Figures 35 and 36 for grades 3 to high school. The item pools for the vertical Scale all tended to measure well over the middle part of the theta distribution. Separation between grades is primarily in the low and high theta ranges.

Establishing the Minimum and Maximum Scale Score

A maximum likelihood procedure will not result in theta estimates for students with perfect or zero scores. Scale scores can be established for these extreme values following a non-maximum likelihood but logical procedure. These minimum and maximum values are called the Lowest Obtainable Theta (LOT) and the Highest Obtainable Theta (HOT). The guidelines for establishing the LOT and HOT values were as follows.

1. The HOT should be high enough so that it does not cause an unnecessary pileup of scale scores at the top of the scale. Likewise, the LOT should be low enough so that it does not cause an unnecessary pileup of scale scores at the bottom part of the scale.
2. The HOT should be low enough so that $CSEM(HOT) < 10 * \text{Min}(CSEMs \text{ for all scale scores})$, where $CSEM$ is the conditional standard error of measurement. The LOT should be high enough so that $CSEM(LOT) < 15 * \text{MIN}(CSEMs \text{ for all scale scores})$.
3. For multiple test levels placed on the same vertical scale, the HOT and LOT values should increase and transition smoothly over levels.

Table 25 provides recommendations for Smarter Balanced for the LOT and HOT values. The LOT and HOT values give the effective range of the ELA/literacy and mathematics scales. The ELA/literacy scale ranges from a value of -4.5941, which is the LOT for grade 3, to the HOT of 3.3392 for high school. In mathematics, the range was from -4.1132 to 4.3804. The means and SDs for theta given in Table 26 reflect the application of these LOT and HOT values.

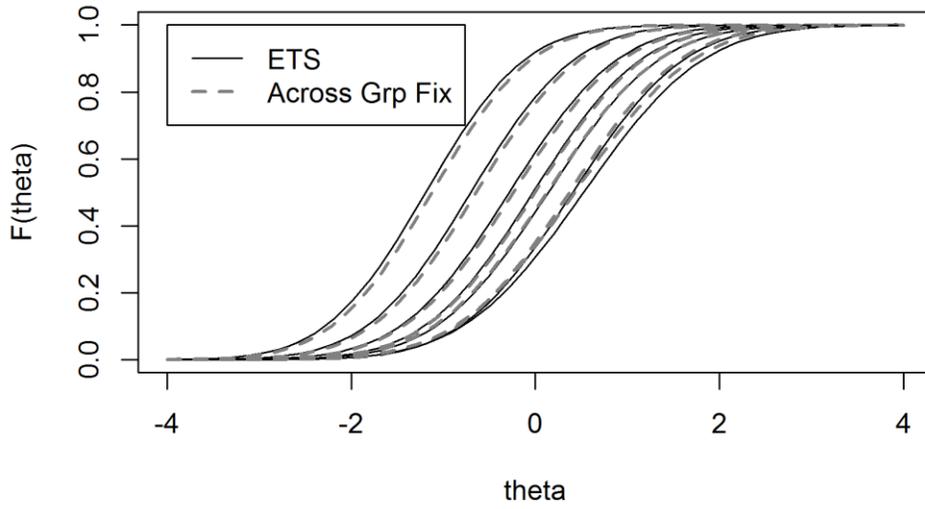
Table 25. Lowest and Highest Obtainable Theta Values and Resulting Theta Scale Summary.

LOT/HOT on Theta Scale						
Grade	LOT	<i>CSEM</i>	HOT	<i>CSEM</i>	Mean	<i>SD</i>
ELA/literacy						
3	-4.5941	1.22	1.3374	0.35	-1.240	1.06
4	-4.3962	1.20	1.8014	0.53	-0.748	1.11
5	-3.5763	1.03	2.2498	0.46	-0.310	1.10
6	-3.4785	1.10	2.5140	0.40	-0.055	1.11
7	-2.9114	0.84	2.7547	0.43	0.114	1.13
8	-2.5677	1.08	3.0430	0.35	0.382	1.13
HS	-2.4375	1.00	3.3392	0.47	0.529	1.19
Mathematics						
3	-4.1132	1.00	1.3335	0.44	-1.285	0.98
4	-3.9204	0.74	1.8191	0.47	-0.708	1.00
5	-3.7276	1.43	2.3290	0.34	-0.345	1.09
6	-3.5348	2.13	2.9455	0.66	-0.131	1.17
7	-3.3420	2.46	3.3238	0.56	-0.060	1.29
8	-3.1492	2.05	3.6254	0.68	0.080	1.36
HS	-2.9564	2.02	4.3804	0.64	0.417	1.47

Cross-validation of Vertical Linking Results

National Center for Research on Evaluation, Standards, and Student Testing (CRESST) was contracted by Smarter Balanced to conduct an independent replication of the Field Test item calibration and vertical linking. These analyses were conducted on the vertical scaling data in which classical item exclusion logic had already been applied. CRESST replicated the within-group concurrent calibration followed by the stepwise application of the Stocking-Lord to perform the vertical linking. They also reported the cross-validation by applying a multigroup approach in which all test levels are calibrated simultaneously. These analyses were implemented using the program **flexMIRT** (Cai, 2013), which implement Bayesian approaches for IRT parameter estimation. There was good agreement between both the CRESST Stocking-Lord and the multigroup approaches with the original ones reported here. Figure 37 shows the cumulative distributions for both the CRESST multigroup approach and the ones implemented for Smarter Balanced. Note the Expected A Posteriori (EAP) scores were computed by CRESST rather than the MLE estimates reported here. Using different methods (i.e., multigroup and Bayesian estimation) had essentially the same outcomes.

CDF of Normal Approx. using EAP Scores: ELA



CDF of Normal Approx. using EAP Scores: Math

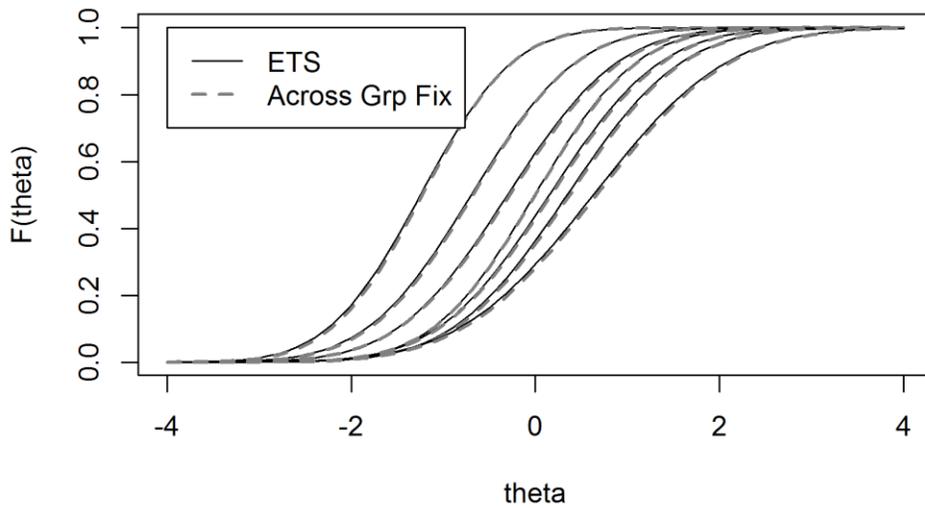


Figure 37. Cross-validation of Vertical Linking Results Comparing cumulative frequency distributions of theta (EAP) for ELA and mathematics obtained from the CRESST cross-validation

Calibration Step for Item Pool

In a second step after the completion of the vertical scaling, all items in the remaining pool were calibrated. This horizontal item-pool calibration involved a much larger number of items and students compared with the initial vertical scaling. It resulted in the final operational item pool at the conclusion of the Field Test. To perform this linking, on-grade, vertical scaling items were administered to students targeted in the horizontal item pool calibration. Using the common items from the vertical scaling step, these on-grade items were linked horizontally onto the scale in each grade using Stocking-Lord test-characteristic-curve methods. Table 26 shows the distribution of observations per student in the item-pool calibration sample after test delivery. Items with fewer than 500 observations were not calibrated. Table 27 presents the mean and standard deviation for the parameter estimates and the number of combined CAT and performance task items. To compare the results of the two respective scaling steps, Figures 38 and 39 show the plots of cumulative theta estimate distributions for the vertical scaling (achievement level setting sample) and the item-pool calibration step. These figures show the outcomes for ELA/literacy and mathematics at each grade. The figures show close agreement for the outcomes from the vertical scaling and item-pool calibration step.

Table 26. Distribution of Student Observations per Item in the Field Test Pool.

Item Response Frequency Percentiles											
Grade	Min	1	5	10	25	Median	75	90	95	99	Max
ELA/literacy											
3	72	270	636	846	1,514	2,438	3,899	6,763	8,123	13,474	24,446
4	80	279	655	957	1,350	2,192	4,171	6,852	8,724	15,945	43,327
5	57	146	658	941	1,296	2,166	4,090	6,740	8,905	12,688	25,518
6	63	242	763	1,128	1,619	2,175	4,202	7,175	11,286	17,193	21,795
7	50	188	556	932	1,284	2,292	4,169	6,735	8,785	17,430	37,351
8	61	253	558	938	1,244	2,057	4,487	7,524	12,702	16,500	20,843
HS	50	73	180	406	849	1,633	3,052	5,635	7,682	13,858	27,229
Mathematics											
3	1,142	1,173	1,251	1,580	1,786	1,914	3,026	4,295	4,838	10,305	14,182
4	913	1,000	1,053	1,220	1,852	2,063	3,782	5,496	6,870	13,121	20,497
5	926	997	1,108	1,258	1,905	2,163	3,769	5,853	8,278	19,590	21,085
6	959	970	987	996	1,331	1,829	2,362	3,993	4,870	11,905	15,702
7	506	910	943	965	1,420	1,940	2,887	3,968	4,656	13,769	14,204
8	618	905	939	957	1,289	1,989	2,765	4,559	5,445	12,964	15,932
HS	164	302	324	339	557	1,066	1,376	2,694	3,241	8,227	30,833

Table 27. Summary of IRT Item Parameter Estimates for the Field Test Item Pool.

<i>a</i> -parameter				<i>b</i> -parameter	
Grade	No. of Items	Mean	SD	Mean	SD
3	896	0.654	0.23	-0.208	1.21
4	856	0.593	0.21	0.259	1.29
5	823	0.613	0.20	0.607	1.23
6	849	0.568	0.22	1.101	1.35
7	875	0.567	0.23	1.333	1.39
8	836	0.555	0.21	1.464	1.43
HS	2,371	0.491	0.18	1.819	1.47
3	1,114	0.851	0.29	-0.759	1.06
4	1,130	0.814	0.29	-0.052	1.03
5	1,043	0.766	0.30	0.669	1.02
6	1,018	0.715	0.26	1.029	1.18
7	942	0.727	0.29	1.670	1.24
8	894	0.626	0.27	2.174	1.41
HS	2,026	0.536	0.26	2.668	1.55

For the item-pool calibration sample, Tables 28 and 29 present the distributions for theta estimates and the conditional standard errors of estimate in a grade and content area using the five-number summary. Table driven methods using sufficient statistics were applied to produce the estimated theta values. The *CSEM* was reciprocal of the inverse of test information for a given student. In this case, the LOT and HOT values were applied for these theta values in each grade and content area.

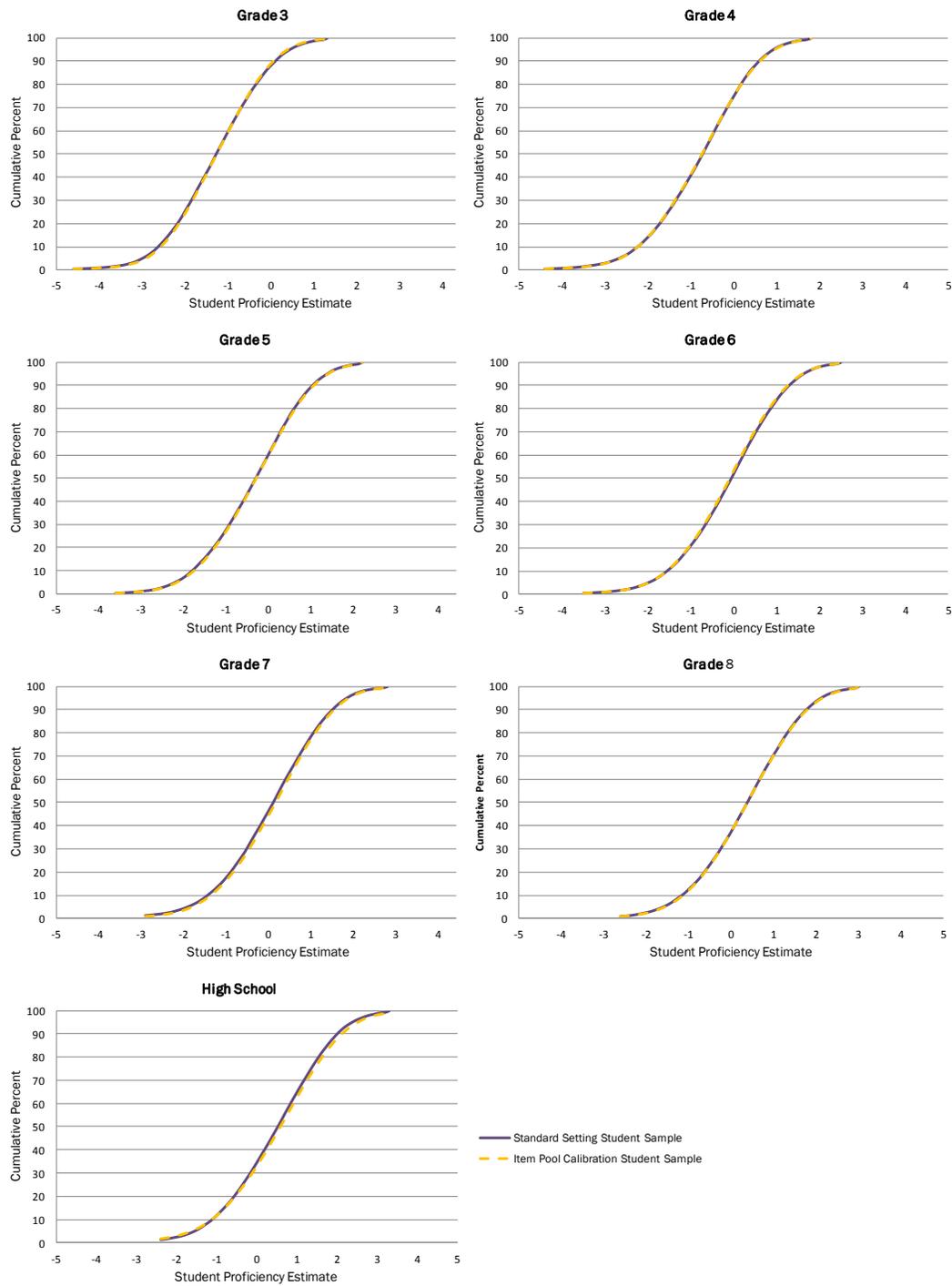


Figure 38. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement Level Setting Sample) and the Item Pool Calibrations Step for ELA/literacy

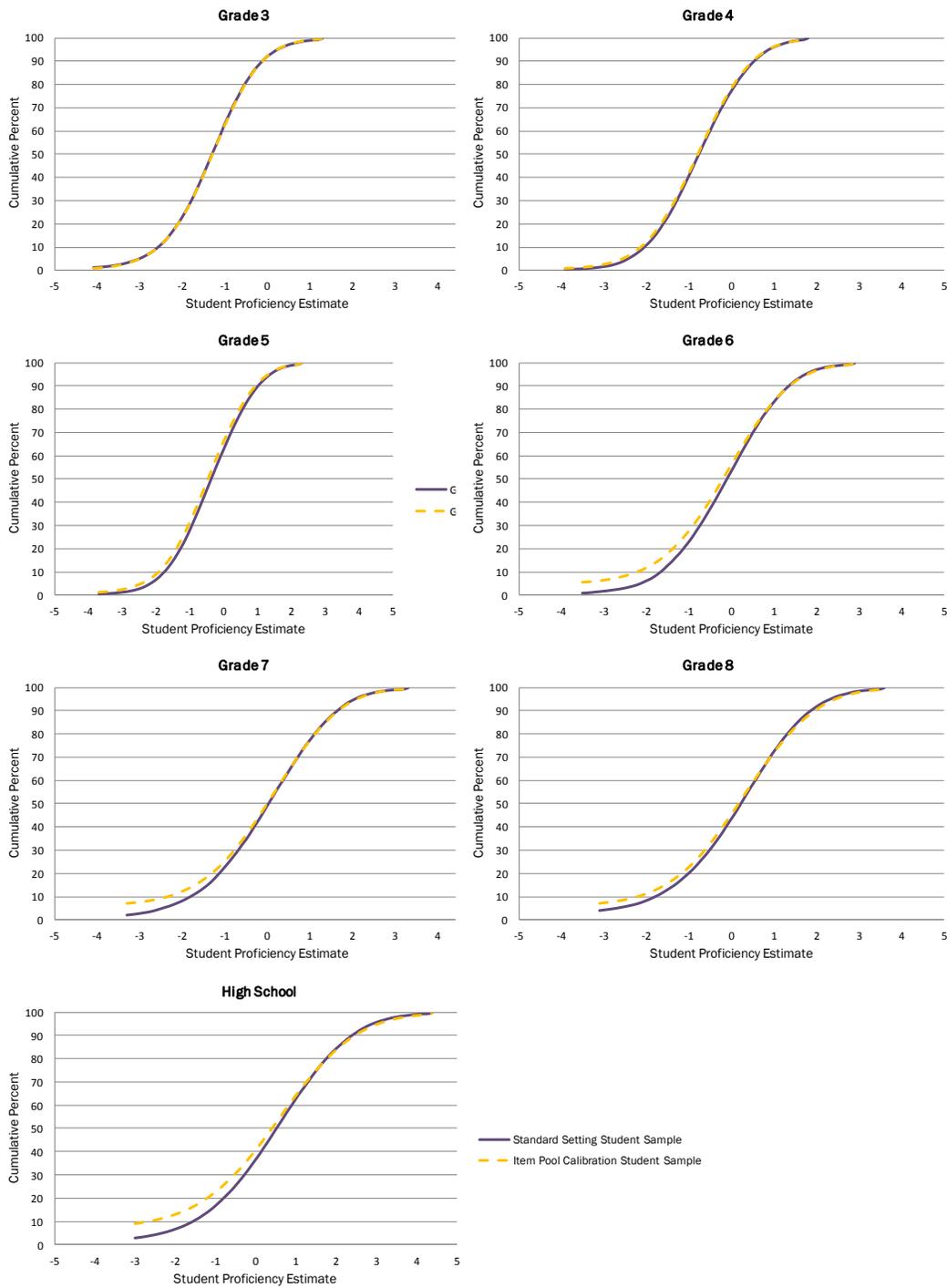


Figure 39. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement Level Setting Sample) and the Item Pool Calibrations Step for Mathematics

Table 28. Distributions of ELA/literacy Theta Estimates and Conditional Standard Error of Measurement.

Grade	3		4		5		6		7		8		HS	
<i>N</i>	83,531		92,595		85,885		90,814		89,332		93,877		228,136	
	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>										
Mean	-1.22	0.38	-0.74	0.43	-0.29	0.40	-0.08	0.40	0.18	0.40	0.41	0.40	0.60	0.49
<i>SD</i>	0.99	0.12	1.07	0.15	1.06	0.12	1.06	0.13	1.08	0.12	1.08	0.11	1.17	0.14
Min	-4.59	0.24	-4.39	0.25	-3.56	0.25	-3.48	0.25	-2.91	0.25	-2.57	0.25	-2.44	0.28
Max	1.34	1.65	1.80	1.90	2.25	1.64	2.51	1.86	2.75	1.62	3.04	1.69	3.34	1.87
10	-2.50	0.28	-2.16	0.30	-1.69	0.30	-1.48	0.30	-1.26	0.30	-1.03	0.31	-0.98	0.36
25	-1.92	0.30	-1.48	0.34	-1.02	0.32	-0.80	0.32	-0.56	0.32	-0.35	0.33	-0.23	0.39
50	-1.22	0.35	-0.69	0.39	-0.25	0.37	-0.04	0.36	0.22	0.36	0.44	0.36	0.64	0.45
75	-0.50	0.41	0.04	0.48	0.49	0.43	0.68	0.43	0.96	0.43	1.20	0.43	1.47	0.54
90	0.08	0.50	0.63	0.59	1.08	0.54	1.29	0.55	1.57	0.54	1.81	0.53	2.13	0.66

Table 29. Distributions of Mathematics Theta Estimates and Conditional Standard Error of Measurement.

Grade	3		4		5		6		7		8		HS	
<i>N</i>	91,325		106,124		105,335		108,667		103,296		102,176		202,115	
	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>	Theta	<i>CSEM</i>
Mean	-1.27	0.33	-0.75	0.37	-0.43	0.44	-0.10	0.56	0.12	0.57	0.31	0.73	0.61	0.92
<i>SD</i>	0.93	0.10	0.99	0.14	1.07	0.19	1.18	0.26	1.24	0.26	1.30	0.37	1.46	0.46
Min	-4.11	0.21	-3.92	0.21	-3.73	0.21	-3.53	0.25	-3.34	0.22	-3.15	0.27	-2.96	0.29
Max	1.33	2.41	1.82	2.00	2.33	2.20	2.94	3.99	3.32	2.49	3.63	4.39	4.38	4.53
10	-2.48	0.25	-2.03	0.26	-1.81	0.28	-1.66	0.33	-1.54	0.31	-1.41	0.39	-1.33	0.48
25	-1.87	0.27	-1.41	0.29	-1.15	0.31	-0.89	0.39	-0.73	0.39	-0.57	0.47	-0.40	0.59
50	-1.24	0.30	-0.73	0.34	-0.42	0.37	-0.07	0.49	0.15	0.51	0.33	0.62	0.62	0.80
75	-0.63	0.35	-0.07	0.41	0.32	0.49	0.73	0.64	1.00	0.68	1.23	0.86	1.63	1.12
90	-0.09	0.44	0.52	0.52	0.95	0.67	1.41	0.86	1.72	0.91	2.00	1.21	2.52	1.52

References

- Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. pp. 395-479. In Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.
- Bock, R. D. & Zimowski, M. F. (1997). Multiple Group IRT. In W.J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433-448). New York: Springer-Verlag.
- Briggs, D. C. & Weeks, J. P. (2009). The Impact of Vertical Scaling Decisions on Growth Interpretations. *Educational Measurement: Issues & Practice*, 28, 3-14.
- Cai, L. (2013). flexMIRT® 2.0: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.
- du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.
- Ercikan, K., Schwarz, R., Julian, M., Burket, G., Weber, M., & Link, V. (1998). Calibration and Scoring of Tests with Multiple-choice and Constructed-Response Item Types. *Journal of Educational Measurement*, 35, 137-155.
- Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling Performance Assessments: A Comparison of One-Parameter and Two-Parameter Partial Credit Models. *Journal of Educational Measurement*, 33, 291-314.
- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C., (1999). *Uncommon Measures: Equivalence and Linkage Among Educational Tests*. Washington, DC: National Academy Press.
- Gu, L., Lall, V. F., Monfils, L., & Jiang, Y. (2010). *Evaluating Anchor Items for Outliers in IRT Common Item Equating: A Review of the Commonly Used Methods and Flagging Criteria*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.
- Haebera, T. (1980). Equating Logistic Ability Scales by Weighted Least Squares Method. *Japanese Psychological Research*, 22, 144-149.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.
- Hanson, B. A., & Beguin, A. A. (1999). *Separate Versus Concurrent Estimation of IRT Parameters in the Common Item Equating Design*. ACT Research Report 99-8. Iowa City, IA: ACT.
- Hanson, B. A., & Beguin, A. A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement*, 26, 3-24.
- Ito, K. Sykes, R. C., & Yao, L. (2008). Concurrent and Separate Grade-Group Linking Procedures for Vertical Scaling. *Applied Measurement in Education*, 21, 187-206.
- Karkee, T., Lewis, D. M., Hoskins, M., Yao, L., & Haug, C. (2003). *Separate Versus Concurrent Calibration Methods in Vertical Scaling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Kim, S. H., & Cohen, A. S. (1998). A Comparison of Linking and Concurrent Calibration Under Item Response Theory. *Applied Psychological Measurement*, 22, 131-143.

- Kim, S. H., & Kolen, M. J. (2004). *STUIRT: A Computer Program for Scale Transformation Under Unidimensional Item Response Theory Models (Version 1.0)*. Iowa Testing Programs, University of Iowa.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating: Methods and Practices*. (2nd ed.). New York, NY: Springer-Verlag.
- Kolen, M. J. (2011). *Issues Associated with Vertical Scales for PARCC Assessments*. PARCC.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.
- Mislevy, R. J. (1987). Recent Developments in Item Response Theory. *Review of Research in Education*, 15, 239-275.
- Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item Analysis and Test Scoring with Binary Logistic Model* (2nd ed.) [Computer program]. Mooresville, IN: Scientific Software.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1: IRT based item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software International, Inc.
- Orlando, M., & Thissen, D. (2003) Further Examination of the Performance of S-X2, an Item Fit index for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, 27(4), 289-98.
- PARPLOT (2009). ETS: Author.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, Norming, and Equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). New York, NY: Macmillan.
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103, 677-680.
- Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, 7, 201-210.
- Stone, C. A., & Zhang, B. (2003). Assessing Goodness of Fit of Item Response Theory Models: A Comparison of Traditional and Alternative Procedures. *Journal of Educational Measurement*, 40, 331-352.
- Stout, W. (1987). A Nonparametric Approach for Assessing Latent Trait Unidimensionality. *Psychometrika*, 52, 589-617.
- Sykes, R. C., & Yen, W. M. (2000). The Scaling of Mixed-Item-Format Tests with the One-Parameter and Two-Parameter Partial Credit. *Journal of Educational Measurement*, 37, 221-244.
- Wainer, H. (Ed.). (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Williams, V. S., Pommerich, M., & Thissen, D. (1998). A Comparison of Developmental Scales Based on Thurstone Methods and Item Response Theory. *Journal of Educational Measurement*, 35, 93-107.
- Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
- Yen, W., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-153). Westport, CT: Praeger Publishers.

- Yen, W. M. (1993). Scaling performance assessments – strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.
- Yen, W. (1981). Using Simulation Results to Choose a Latent Trait Model. *Applied Psychological Measurement, 5*, 245-262.
- Yen, W. M. (1986). The Choice of Scale for Educational Measurement: An IRT Perspective. *Journal of Educational Measurement, 23*, 299-325.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. D. (1997). *BILOG-MG: Multiple Group Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software.