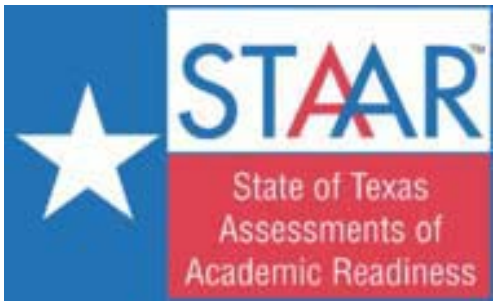


State of Texas Assessments of Academic Readiness (STAAR™)

Vertical Scale Technical Report

June 2013

Table of Contents

Chapter 1: The State of Texas Assessments of Academic Readiness (STAAR™)

This chapter provides an overview of the STAAR program and includes high-level descriptions of the following:

- STAAR Vertical Scale
- Goals of the STAAR Program
- STAAR Test Design

STAAR Vertical Scale

Under Texas Education Code (TEC) §39.036, the Texas Education Agency (TEA) is required to develop a vertical scale for assessing student performance in grades 3–8 for reading and mathematics. A vertical scale is a scale score system that allows for direct comparison of student test scores across grade levels within a content area. Vertical scaling refers to the process of placing test scores that measure similar content areas but at different grade levels onto a common scale. A vertical scale was developed for the following grades and subjects:

- STAAR English grades 3–8 mathematics
- STAAR English grades 3–8 reading
- STAAR Spanish grades 3–5 reading

Although there is a Spanish version of STAAR mathematics assessments in grades 3–5, a separate vertical scale was not developed because the same scale is used for both language versions. Use of the same scale is possible because Spanish mathematics items are transadapted from the English items. Spanish reading passages and items are uniquely developed to maintain the authenticity of the Spanish assessment. Therefore, Spanish reading items must be field-tested and placed on a unique Spanish reading scale score system. As a result, a separate vertical scaling study for Spanish reading was conducted. STAAR assessments are also available for science, social studies, and writing at the elementary and middle school levels. However, vertical scales are not available for these subjects.

The following sections provide a general introduction to the STAAR assessment program goals and test design considerations that influenced decisions made throughout the development of the STAAR vertical scales. These sections describe how the STAAR program is vertically aligned across grade levels through the curriculum standards, content standards, and performance standards. The inherent vertical alignment of the STAAR assessments provides a strong basis for the implementation of a vertical scale. In order to implement a vertical scale, research studies were needed to determine differences in difficulty across grade levels. This report provides a summary of the data collection design, analysis methodology, results, and implementation of the STAAR 3–8 reading and mathematics vertical scale study.

Goals of the STAAR Program

The 80th and 81st sessions of the Texas Legislature called for a new state assessment program to replace the Texas Assessment of Knowledge and Skills (TAKS). One of the state's goals in developing STAAR was that Texas should be among the top 10 states for graduating college- and career-ready students by the 2019–2020 school year.

Toward this end, TEA, in collaboration with the Texas Higher Education Coordinating Board (THECB) and Texas educators, has developed STAAR to be a more rigorous assessment. STAAR is based on a new assessment model that includes the following:

- Performance expectations for STAAR were established so that graduating students would receive feedback about their level of postsecondary readiness in STAAR Algebra II and English III assessments and the degree to which they were on track toward postsecondary readiness in preceding assessments.
- The STAAR program was designed to be a comprehensive system, with curriculum and performance standards aligned with and linked from high school back to elementary and middle school (grades 3–8) and projecting forward to postsecondary readiness. Figure 1 provides a visual representation of this goal for the STAAR program.

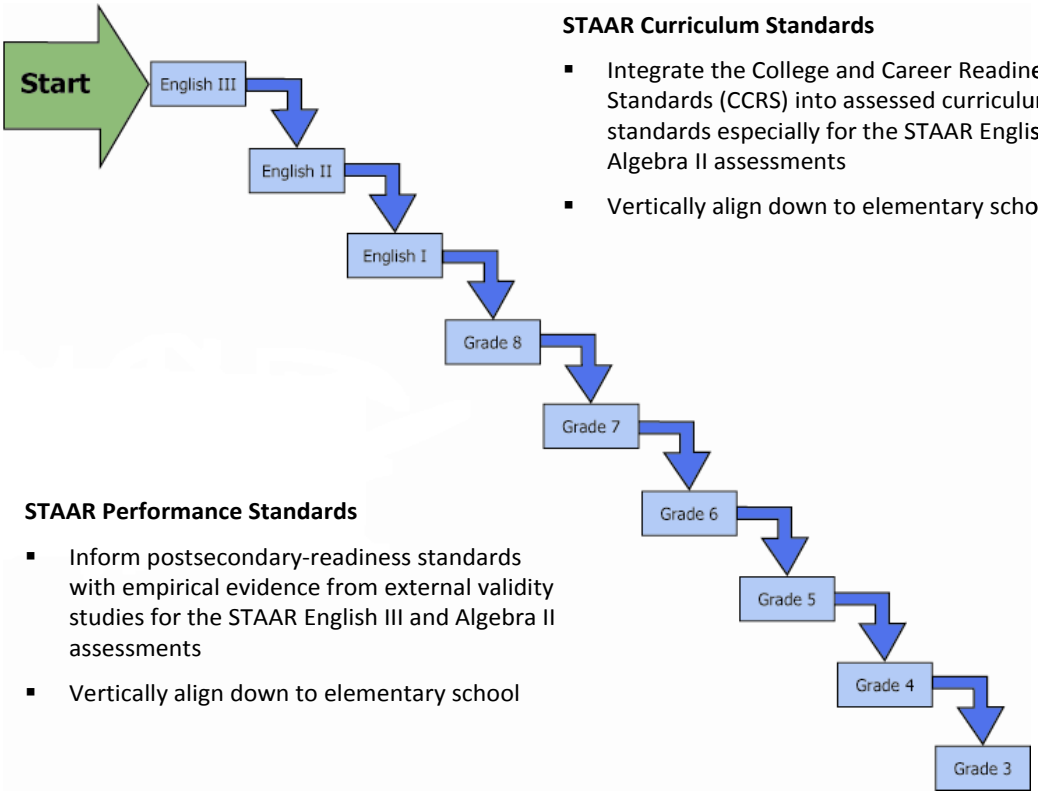


Figure 1. Vertical Alignment of Curriculum and Performance Standards for the STAAR Program

The sections that follow provide a high-level description of how the curriculum is assessed with STAAR in order to meet the goals and requirements of the new assessment program.

STAAR Test Design

The curriculum assessed on STAAR is the state-mandated curriculum standards, the Texas Essential Knowledge and Skills (TEKS). These standards are designed to prepare students to succeed in postsecondary opportunities and to compete globally. However, consistent with a growing national consensus regarding the need to provide a more clearly articulated K–16 education program, STAAR focuses on fewer skills and addresses those skills in a deeper manner. By focusing on the TEKS that are most critical to assess, STAAR measures the academic performance of students as they progress from elementary to middle to high school.

Based on educator committee recommendations for each grade or course, TEA has identified a set of knowledge and skills from the TEKS that are eligible to be assessed. One subset of the TEKS, called readiness standards, is emphasized on the assessments. Other knowledge and skills are considered supporting standards and are assessed, although not emphasized.

Readiness standards have the following characteristics:

- They are essential for success in the current grade level or course.
- They are important for preparedness for the next grade level or course.
- They support postsecondary readiness.
- They necessitate in-depth instruction.
- They address broad and deep ideas.

Supporting standards have the following characteristics:

- Although introduced in the current grade or course, they may be emphasized in a subsequent grade or course.
- Although reinforced in the current grade or course, they may be emphasized in a previous grade or course.
- They play a role in preparing students for the next grade or course but not one that is central.
- They address more narrowly defined ideas.

The STAAR assessment blueprints are designed so that a larger number of test items measure student expectations designated as readiness standards. The readiness standards emphasize the vertical alignment of the curriculum and carry this forward to the test design.

TEA has also implemented a number of changes in the STAAR test design that serve to assess knowledge and skills in a deeper way.

- Tests contain a greater number of items that have a higher cognitive complexity level.
- Questions are developed to more closely match the cognitive complexity level evident in the TEKS.
- In reading, greater emphasis is given to critical analysis than to literal understanding.
- In mathematics, process skills are assessed in context, not in isolation, which allows for a more integrated and authentic assessment of these content areas.
- In mathematics, the number of open-ended (griddable) questions has increased to allow students more opportunity to derive an answer independently.

In addition to the changes in test design, the TEA Curriculum division implemented the revised TEKS for the reading content area which were assessed for the first time through STAAR assessments in spring 2012 (http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147506272&libID=2147506265).

The joint *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999; Standards 3.25 and 3.26) recommends evaluating the impact of score interpretations when revisions are made to an assessment. As a result of the notable differences between the STAAR and TAKS programs, a new vertical scale study was conducted for the STAAR 3–8 mathematics and reading programs. The following chapters discuss the data collection design, selection of the vertical scale items to represent the blueprint and readiness/supporting standards, methodology for establishing the vertical scale, the results, an evaluation of the results, and the implementation of the STAAR vertical scale study.

Chapter 2: Methodology

This chapter provides the methodology for the STAAR 3–8 vertical scale study and includes the following:

- Data Collection Design
- Selection of the Vertical Scale Items
- Analysis Method

Data Collection Design

To avoid burdening schools by requiring participation in a separate vertical scale study, TEA collected data for the vertical scale study during the first operational administration of STAAR in spring 2012. Vertical scale items were embedded in the operational test forms in field-test positions. Students' performance on base-test items determine students' scores, whereas items in field-test positions do not count towards students' scores. In general, field-test positions are used to collect data on new items to determine if the items meet statistical criteria for use on a future test as a base-test item. For the vertical scale study, field-test positions were used to place off-grade-level items onto on-grade-level test forms within a content area. Using the embedded field-test positions is desirable because students have no knowledge of whether an item is a base-test item or vertical scale item. Additionally, including vertical scale items in embedded field-test positions allows cross-grade item position effects to be minimized because the location of field-test blocks is similar across grade levels.

The vertical scale items are referenced in several ways depending on the grade level for the test form and, therefore, the grade level of the students taking the test. *On-grade-level items* are included in a test form that matches their grade level. *Off-grade-level items* are included in an adjacent grade-level test form that is above or below their grade level. The off-grade-level vertical scale items are further defined based on whether the item is from an upper grade-level or a lower grade-level compared to the grade level of the test form.

- On-grade-level item – a vertical scale item from the same grade level
- Off-grade-level item – a vertical scale item from an adjacent grade level
    - Upper-grade-level item – an off-grade-level item from an upper grade level
    - Lower-grade-level item – an off-grade-level item from a lower grade level

Figure 2 illustrates the four vertical scale item categories for a grade 4 test. In the figure, the grade 4 operational test is illustrated at the top of the figure denoting the base-test items and the field-test items. For the purpose of illustration, the field-test items are depicted in the middle of the base-test form. Below the figure of the test are eight boxes each representing a set of vertical scale items that would be placed in the field-test positions in the test. The two boxes with the number 3 represent grade 3 vertical scale items included in the grade 4 test as lower-grade-level items. The two boxes with the number 5 represent grade 5 vertical scale

items included in the grade 4 test as upper-grade-level items. The boxes denoted 3 or 5 represent the off-grade-level items that are administered to grade 4 students. The four boxes with the number 4 are grade 4 vertical scale items, referred to as on-grade-level items. Any individual student would only see the base-test items plus one of these eight sets of vertical scale items.

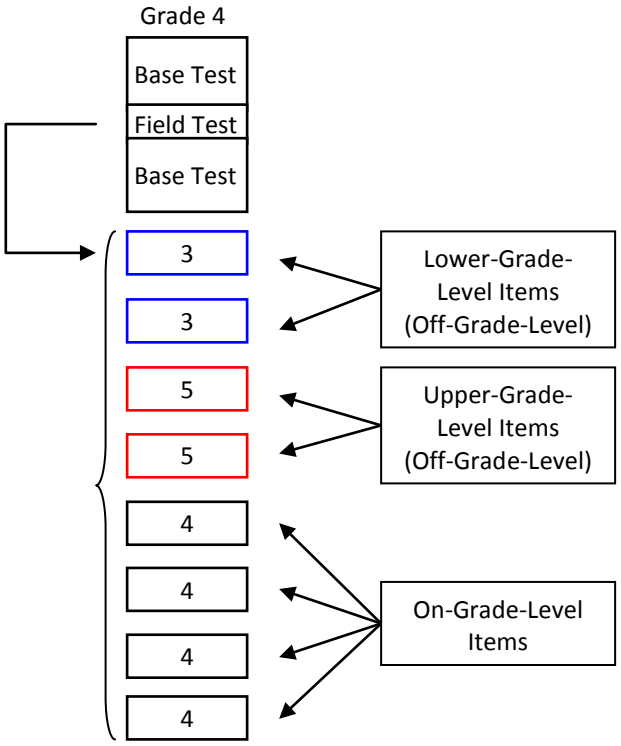


Figure 2. Example Referencing the Types of Vertical Scale Items

The data collection design is a common-item non-equivalent groups design in which students in adjacent grade levels respond to the same items, thereby allowing direct comparison of item difficulties (Kolen & Brennan, 2004). This design allows items to be placed on the same scale. The vertical scale common items between adjacent grade levels determine the relationship between tests in adjacent grades for reading and mathematics. This design requires designating either four or eight of the regular field-test forms for the vertical scale study. Two forms are assigned per grade for each pair of adjacent grades. For grades 4–7, there are both upper and lower adjacent grade levels resulting in a total of four on-grade-level forms and four off-grade-level forms. For grades 3 and 8, there is only one adjacent grade level resulting in two on-grade-level forms and two off-grade-level forms.

The data collection design used with STAAR English 3–8 reading and mathematics assessments is presented in Figure 3. The STAAR Spanish 3–5 reading data collection design is presented in Figure 4. The illustration shows the number of vertical scale forms for each grade and designates the grade level for each form. The vertical scale forms denoted with horizontal

arrows between adjacent grades represent the same set of vertical scale items on both grade-level tests. For example, the grade 4 test has two forms with lower-grade-level vertical scale items from grade 3. These items are also included in the forms for the grade 3 test.

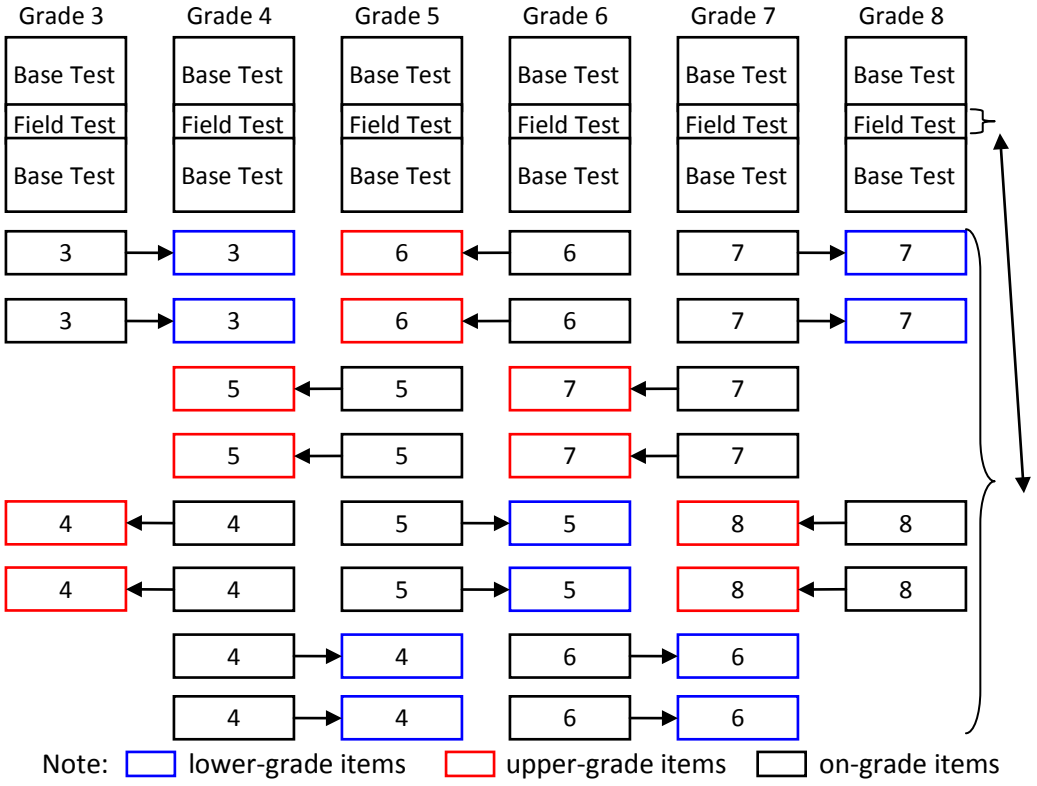


Figure 3. STAAR 3–8 English Reading and Mathematics Vertical Scale Data Collection Design

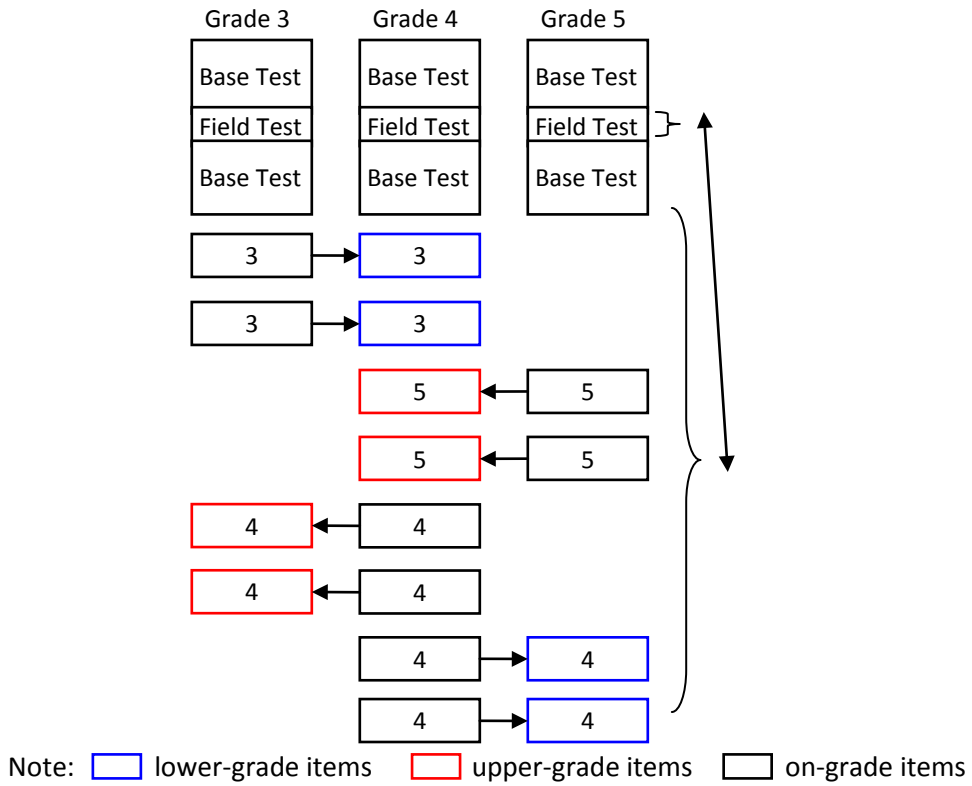


Figure 4. STAAR 3–5 Spanish Reading Vertical Scale Data Collection Design

Each reading test form contained eight field-test positions for vertical scale items and each mathematics test form contained seven field-test positions for vertical scale items. Table 1 lists the number of base-test items, lower-grade-level vertical scale items, and upper-grade-level vertical scale items per grade for STAAR English 3–8 mathematics and reading and STAAR Spanish 3–5 reading.

For some grades and subjects, vertical scale items were duplicated across more than one form due to constraints of the available items in the item bank. As a result, the number of vertical scale items listed in Table 1 does not exactly correspond to the total number of vertical scale items possible per form for some grades and subjects. For example, the shaded cells in Table 1 shows the number of on-grade vertical scale items for STAAR English grade 4 reading is 28 unique items rather than 32 (four forms with eight vertical scale positions). This issue arose for the reading tests as selection of items was constrained by the availability of items associated with each reading passage. In this example, the 28 grade 4 items are also on the adjacent grades: 16 are on the grade 3 reading test and 12 are on the grade 5 reading test. A second example from Table 1 is the STAAR grade 4 mathematics vertical scale items. There are 27 unique vertical scale items and one item is included on the lower-grade level and the same item is included on the upper-grade level (see shaded cells in Table 1).

Table 1. Number of Base-Test Items and Vertical Scale Items by Grade and Subject

		Grade					
Subject	Item Type	3	4	5	6	7	8
Mathematics	Base Test	46	48	50	52	54	56
	Lower-Grade Vertical Scale	--	14	14	14	14	14
	On-Grade Vertical Scale	14	27	28	28	27	14
	Upper-Grade Vertical Scale	14	14	14	14	14	--
English Reading	Base Test	40	44	46	48	50	52
	Lower-Grade Vertical Scale	--	16	12	12	16	16
	On-Grade Vertical Scale	16	28	28	32	32	16
	Upper-Grade Vertical Scale	16	16	15	16	16	--
Spanish Reading	Base Test	40	44	46	--	--	--
	Lower-Grade Vertical Scale	--	16	12	--	--	--
	On-Grade Vertical Scale	16	27	16	--	--	--
	Upper-Grade Vertical Scale	15	16	--	--	--	--

Note: There are eight field-test positions per form. However, for mathematics the griddable item is not included in the vertical scale item set resulting in seven field-test positions per form.

Selection of Vertical Scale Items

Vertical scale items are items previously field tested that met the content and psychometric guidelines provided for the vertical scale study. The guidelines for selecting the vertical scale items was reviewed by a Texas Technical Advisory Committee member in April 2011 (see Appendix 1). The selection process for vertical scale items consisted of content specialists selecting items based on the content guidelines after which psychometricians evaluated the items for psychometric properties. Since the items were embedded in field-test positions, the vertical scale items were evaluated against the base-test items to confirm that the vertical scale items did not clue students' responses to base-test items. The following describes the content and psychometric guidelines for selecting the vertical scale items.

Content Guidelines

In general, the on-grade-level vertical scale items represent the test blueprint for the grade level. For example, the 28 on-grade-level vertical scale items for STAAR grade 4 reading listed in Table 1 represent the grade 4 test blueprint with respect to reporting categories and readiness/supporting standards. On-grade level items were selected first to match the test blueprint and then evaluated for degree of content overlap with each adjacent grade (either upper or lower). When selecting upper-grade-level vertical scale items to move to a lower grade, special attention was paid to make sure the items came from a content area where lower grade students would have had exposure to the topic.

The STAAR reading assessments are passage-based; as a result, the vertical scale items were associated with one passage per form. The vertical scale passages were selected to represent the types of passages included on the grade level base test, to the extent possible. The

passages with upper-grade-level vertical scale items were selected to be as close as possible to the on-grade-level requirements in characteristics such as word counts.

The STAAR mathematics assessments include griddable items in which students entered the actual value rather than selecting from a list of answer choices. While it would be desirable to include these griddable items in the development of the vertical scale, the variation in the format of the grids on the answer documents across grade levels made it problematic to include these items without altering either the items themselves or creating specialized answer documents for study purposes. For example, Figure 5 illustrates that the griddable items at grades 4–5 require students to grid up to three whole number values with respect to a decimal point whereas the griddable items at grades 6–8 require students to grid up to four whole numbers and two decimal values. It would, therefore, not be possible to authentically represent a grade 6 griddable item embedded within the grade 5 test.

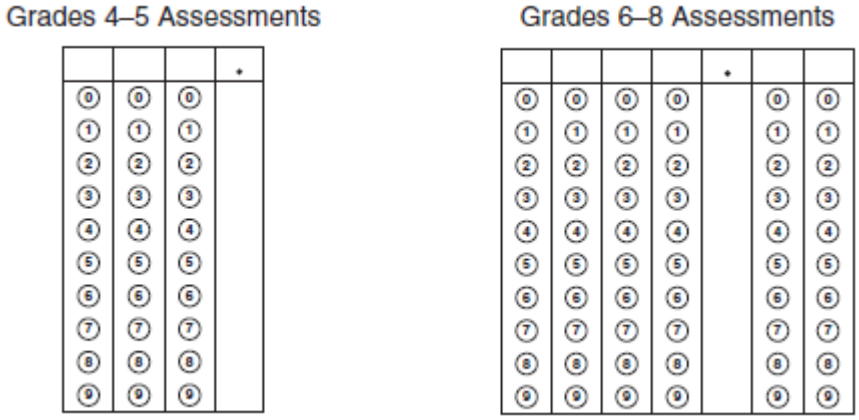


Figure 5. Example STAAR Mathematics Griddable Item Answer Document

Since the STAAR Spanish mathematics assessments for grades 3–5 use the same scale score system as the STAAR English mathematics assessments for grades 3–5, additional criteria were applied to the vertical scale items for these grades. The vertical scale items for STAAR English mathematics were selected only if the items could be transadapted into Spanish.

Tables 2 and 3 list the percent of mathematics items for the base-test (base) and the on-grade level vertical scale items (VS) by grade level for the reporting categories and readiness/supporting standards, respectively. The content representation by reporting category of the on-grade vertical scale items were similar (within 8% difference) to the content representation of the base-test items for the same grade level. The readiness and supporting standards were fairly similar (within 16% difference) to the base-test representation. All the grade levels, except grade 4 mathematics reported similar or higher percentages of items measuring the readiness standards compared to the supporting standards. The available items in the item bank and the degree of content overlap with grades 3 and 5 resulted in more supporting items rather than readiness items for grade 4 mathematics.

Table 2. STAAR 3–8 Mathematics Percent of Base-Test and On-Grade Vertical Scale Items by Reporting Category

Reporting Category	Grade Levels											
	3		4		5		6		7		8	
	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS
1	33%	29%	35%	33%	36%	36%	31%	36%	24%	22%	20%	21%
2	17%	21%	13%	15%	12%	14%	23%	18%	24%	30%	25%	29%
3	20%	21%	25%	22%	14%	14%	15%	14%	19%	19%	14%	21%
4	17%	14%	17%	15%	16%	18%	15%	18%	15%	15%	23%	14%
5	13%	14%	10%	15%	22%	18%	15%	14%	19%	15%	18%	14%
Total Items	46	14	48	27	50	28	52	28	54	27	56	14

Note: Percentages may not add to 100% due to rounding.

Table 3. STAAR 3–8 Mathematics Percent of Base-Test and On-Grade Vertical Scale Items by Readiness and Supporting Standards

Standards	Grade Levels											
	3		4		5		6		7		8	
	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS
Readiness	61%	64%	60%	44%	62%	50%	62%	50%	65%	59%	63%	50%
Supporting	39%	36%	40%	56%	38%	50%	38%	50%	35%	41%	38%	50%
Total Items	46	14	48	27	50	28	52	28	54	27	56	14

Note: Percentages may not add to 100% due to rounding.

Tables 4 and 5 list the percent of reading items for the base-test (base) and the on-grade level vertical scale items (VS) by grade level for the reporting categories and readiness/supporting standards, respectively. The content representation by reporting category of the on-grade vertical scale items was fairly similar (within 19% difference) to the content representation of the base-test items for the same grade level. The readiness and supporting standards were somewhat similar (within 32% difference) to the base-test representation except for grades 3 and 8. All the grade levels reported higher percentages of the readiness standards compared to the supporting standards. For grades 3 and 8 almost the entire set of vertical scale items were readiness items. The differences in percentages for readiness and supporting standards were a result of limitations in the item bank and dependency on passage-based items.

Table 4. STAAR 3–8 Reading Percent of Base-Test and On-Grade Vertical Scale Items by Reporting Category

Reporting Category	Grade Levels											
	3		4		5		6		7		8	
	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS
1	15%	25%	23%	14%	22%	11%	21%	23%	20%	13%	19%	19%
2	45%	38%	41%	50%	41%	39%	42%	58%	42%	41%	42%	44%
3	40%	38%	36%	36%	37%	50%	38%	19%	38%	47%	38%	38%
Total Items	40	16	44	28	46	28	48	31	50	32	52	16

Note: Percentages may not add to 100% due to rounding.

Table 5. STAAR 3–8 Reading Percent of Base-Test and On-Grade Vertical Scale Items by Readiness and Supporting Standards

Standards	Grade Levels											
	3		4		5		6		7		8	
	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS	Base	VS
Readiness	63%	94%	61%	89%	70%	50%	67%	81%	70%	72%	65%	94%
Supporting	38%	6%	39%	11%	30%	50%	33%	19%	30%	28%	35%	6%
Total Items	40	16	44	28	46	28	48	31	50	32	52	16

Note: Percentages may not add to 100% due to rounding.

Tables 6 and 7 list the percent of Spanish reading items for the base-test (base) and the on-grade level vertical scale items (VS) by grade level for the reporting categories and readiness/supporting standards, respectively. The content representation by reporting category of the on-grade vertical scale items were similar (within 10% difference) to the content representation of the base-test items for the same grade level. For grades 3 and 5 the percent of readiness and supporting items were almost identical between the base test and vertical scale items. For grade 4, almost the entire set of vertical scale items were readiness items. this difference was a result of limitations in the item bank and dependency on passage-based items.

Table 6. STAAR Spanish 3–5 Reading Percent of Base-Test and On-Grade Vertical Scale Items by Reporting Category

Reporting Category	Grade Levels					
	3		4		5	
	Base	VS	Base	VS	Base	VS
1	15%	25%	23%	22%	22%	19%
2	45%	38%	41%	44%	41%	38%
3	40%	38%	36%	33%	37%	44%
Total Items	40	16	44	27	46	16

Note: Percentages may not add to 100% due to rounding.

Table 7. STAAR Spanish 3–5 Reading Percent of Base-Test and On-Grade Vertical Scale Items by Readiness and Supporting Standards

	Grade Levels					
	3		4		5	
Standards	Base	VS	Base	VS	Base	VS
Readiness	63%	63%	61%	93%	63%	63%
Supporting	38%	38%	39%	7%	37%	38%
Total Items	40	16	44	27	46	16

Note: Percentages may not add to 100% due to rounding.

Psychometric Guidelines

The vertical scale item-selection guidelines included psychometric properties of the individual items with respect to the other on-grade-level items. Each item was selected such that it met classical test theory and item response theory criteria. The following lists the psychometric criteria for individual items:

- Select items field-tested within the past three years for mathematics. For reading items, the implementation of the new reading curriculum in 2010–2011 resulted in all items field-tested within two years.
- Avoid extremely easy or difficult items such that classical item difficulty (p-value) is within a range of 0.20-0.90.
- Point biserials should be greater than or equal to 0.20.
- Rasch item fit should be between 0.80 to 1.20.

The on-grade-level items were evaluated together to make sure there was a range of Rasch item difficulties and that the overall Rasch difficulty level of the on-grade-level vertical scale items were similar to the overall difficulty of the base-test items. Tables 8 and 9 provide the mean and standard deviation of the Rasch item difficulties for the base-test items and the on-grade-level vertical scale items for STAAR English mathematics and reading and STAAR Spanish reading, respectively. For STAAR 3–8 mathematics, the mean and standard deviations of the Rasch item difficulties for the base-test and vertical scale items were fairly similar (differences were within 0.45 for the mean and within 0.30 for the standard deviation).

For STAAR English 3–8 reading, the mean and standard deviations of the Rasch item difficulties for the base-test and vertical scale items were similar (differences were within 0.20 for the mean and within 0.51 for the standard deviation). In general, the standard deviations for the vertical scale items were smaller than the base-test items.

Table 8. Summary Statistics for Rasch Item Difficulties for STAAR 3–8 Mathematics and Reading

		Mathematics			Reading		
Grade	Item Type	N	Mean	Standard Deviation	N	Mean	Standard Deviation
3	Base-Test	46	1.14	0.80	40	0.26	0.87
	On-Grade Vertical Scale	14	1.04	0.62	16	0.46	0.36
4	Base-Test	48	1.19	0.79	44	0.35	0.80
	On-Grade Vertical Scale	27	0.95	0.49	28	0.35	0.74
5	Base-Test	50	1.09	0.68	46	0.25	0.87
	On-Grade Vertical Scale	28	0.74	0.77	28	0.21	0.85
6	Base-Test	52	1.12	0.66	48	0.30	0.76
	On-Grade Vertical Scale	28	0.67	0.81	31	0.21	1.01
7	Base-Test	54	0.87	0.65	50	0.34	0.63
	On-Grade Vertical Scale	27	0.89	0.69	32	0.31	0.64
8	Base-Test	56	0.95	0.60	52	0.33	0.77
	On-Grade Vertical Scale	14	0.54	0.67	16	0.19	0.62

For STAAR Spanish 3–5 reading, the mean and standard deviations of the Rasch item difficulties for the base-test and vertical scale items were similar (differences were within 0.11 for the mean and within 0.17 for the standard deviation). In general, the standard deviations for the vertical scale items were slightly larger than the base-test items except for grade 5.

Table 9. Summary Statistics for Rasch Item Difficulties for STAAR Spanish 3–5 Reading

		Spanish Reading		
Grade	Item Type	N	Mean	Standard Deviation
3	Base-Test	40	0.30	0.66
	On-Grade Vertical Scale	16	0.23	0.75
4	Base-Test	44	0.33	0.67
	On-Grade Vertical Scale	27	0.22	0.77
5	Base-Test	46	0.29	0.75
	On-Grade Vertical Scale	16	0.33	0.58

Analysis Method

When the data-collection design is based on common-item non-equivalent groups and items from one test are embedded in an adjacent grade-level test, item response theory places test items and measures of student proficiency on the same scale. The relationship between the adjacent grade-level tests is determined based on the underlying item response theory scale.

The STAAR assessments are scaled and equated using an item response theory model known as the Rasch Partial-Credit Model (RPCM) to place test items and measures of student proficiency on the same scale across assessments. The RPCM is an extension of the Rasch one-parameter Item Response Theory model attributed to Georg Rasch (1966), as extended by Wright and Stone (1979), Masters (1982), Wright and Masters (1982), and Linacre (2001). The RPCM maintains a one-to-one relationship between Rasch-based performance estimates (θ), scale scores, and raw scores, meaning each raw score is associated with a unique scale score.

The RPCM is defined by the following mathematical measurement model where, for a given item/prompt involving *m + 1* score categories, the probability of person *n* scoring *x* on item/prompt *i* is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}, \qquad \mathrm{x} = 0, 1, \ldots, \mathrm{m_i} \tag{1}$$

The RPCM provides the probability of a student scoring *x* on the *m* steps of item/prompt *i* as a function of the student's Rasch-based performance estimates, θ_n, and the step difficulties, δ_{ij}, of the *m* steps in prompt *i* (Refer to Masters, 1982, for an example). Note that for multiple-choice and griddable items, there are only two score categories: (a) 0 for an incorrect response and (b) 1 for a correct response, in which case the RPCM reduces to the standard Rasch one-parameter IRT model, and the resulting single-step difficulty is more properly referred to as a Rasch item difficulty. The method of estimating the Rasch item difficulties is referred to as a calibration.

Vertical scaling using the RPCM occurs in three stages. In the first stage, all items administered (both on- and off-grade) to students at a given grade level are brought onto the same RPCM scale. This can be thought of as scaling "down the columns" in Figure 3 using only student data for a single grade level. In the second stage, the values of the off-grade-level items are compared "across the columns" in a pairwise fashion and these differences are used to bring adjacent grade-level Rasch scales onto the same Rasch scale using data from students in different grade levels. This process is repeated for each set of adjacent grade-level Rasch scales. In the third stage, differences in adjacent grade-level scales are "aggregated" to create a difference between each grade and the anchor grade. The anchor grade is the grade level that will define the Rasch vertical scale across the multiple scales. The vertical scale constants between grades are aggregated to the anchor grade (in this case grade 8) through the other grade levels for the vertical scale (for example the difference between grade 4 and grade 8) such that at the end of the third stage, all grades are on the same Rasch scale—"the vertical scale."

Stage 1 ("Down the Columns")

The RPCM placed all STAAR items within a grade on a common Rasch scale through a two-step calibration process in Stage 1. The first step calibrated only the Rasch item difficulties for the base-test items within a grade level together using all available student data. The second calibration step estimated Rasch item difficulties for the base-test items, on-grade-level vertical scale items, and off-grade-level vertical scale items through an incomplete data matrix (IDM) separately for each grade level. Figure 6 provides an illustration for the two-step process for STAAR grade 4 reading calibrations. The grade 4 base-test items were calibrated together during Step 1. Then the grade 4 base-test items, grade 4 on-grade-level items, grade 3 lower-grade-level items, and the grade 5 upper-grade level items were calibrated together during Step 2. This process is conducted for each grade within a content area.

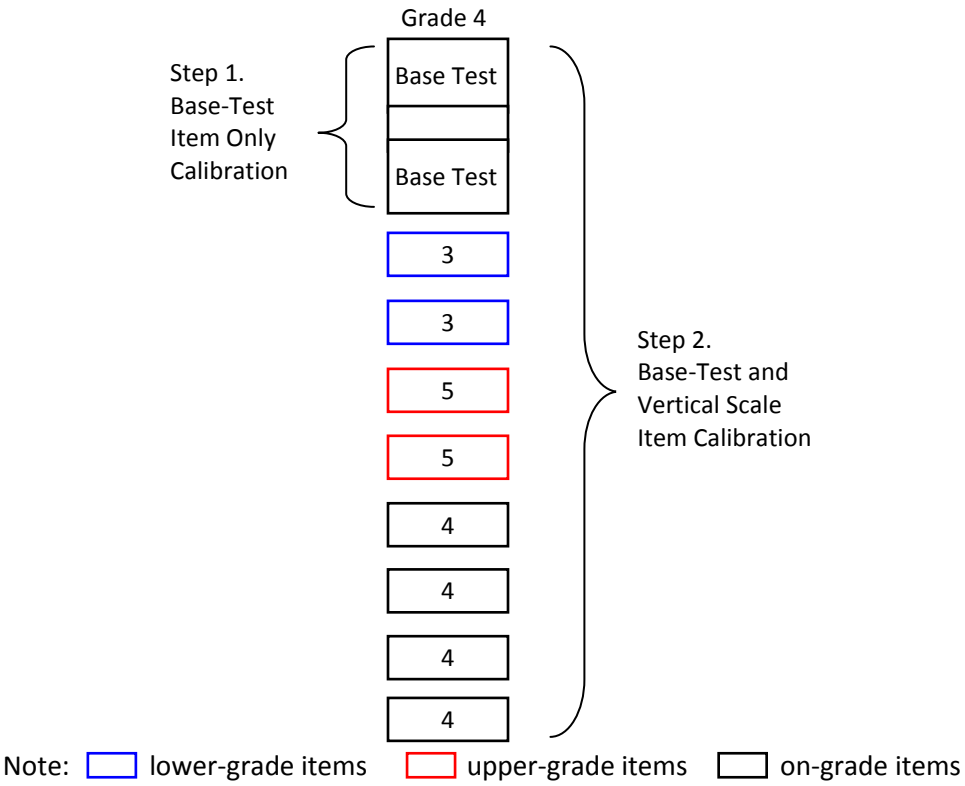


Figure 6. Illustration of STAAR English Grade 4 Reading Step 1 and 2 Calibration "Down the Columns"

Figure 7 illustrates the data structure for the STAAR English grade 4 reading calibrations for Steps 1 and 2. For Step 1, all the base-test items are administered to all the grade 4 students in the sample; therefore, the calibration includes all students who were administered Forms 1 to N, where N is the total number of test forms. For Step 2, the student data for all the base-test items (Forms 1 to N) are included. In addition, the vertical scale items for the eight vertical scale forms are also included in the calibration. Therefore students that were administered a vertical

scale form provided response data for the base-test items and the vertical scale items, resulting in an IDM.

For each grade and subject, the IDM enabled the calibration of all the vertical scale items across all the vertical scale forms in a single step. In addition, the IDM allowed all data for the vertical scale items duplicated across vertical scale forms to be combined. This was done separately by grade and subject resulting in six separate calibrations for STAAR mathematics, six separate calibrations for STAAR English reading and three separate calibrations for STAAR Spanish reading.

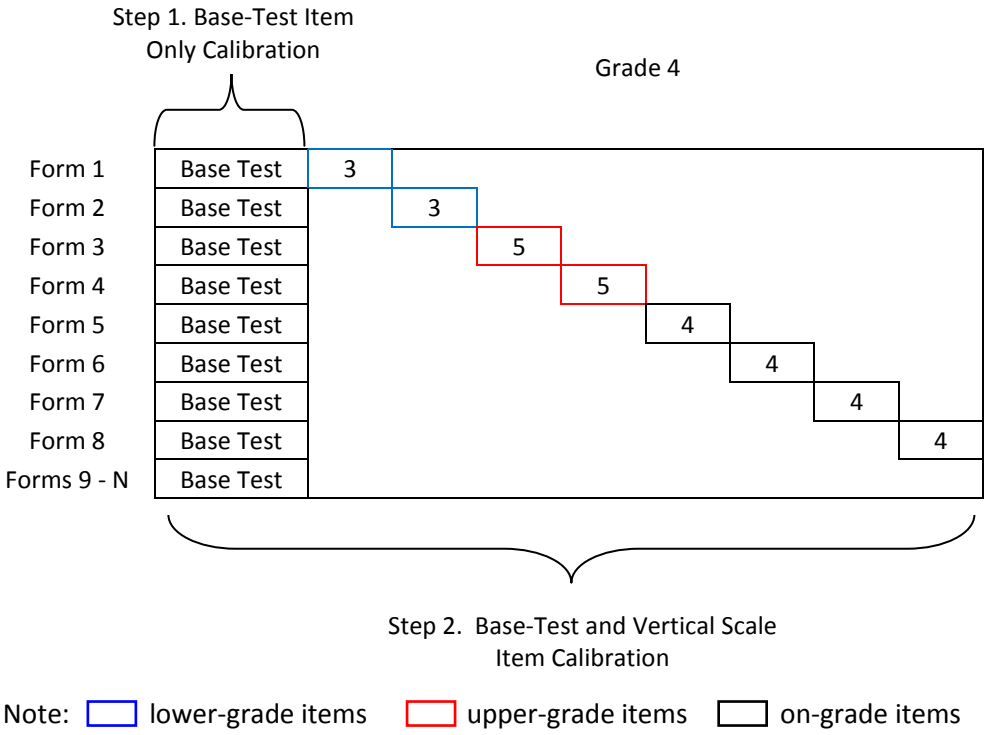


Figure 7. Illustration of STAAR English Grade 4 Reading Data Structure for Step 1 and 2 Calibration

The two-step calibration process allows for a Rasch scale to first be defined with only base-test items. Then a second Rasch scale can be defined with both the base-test and the vertical scale items. The influence of the vertical scale items can then be observed as the difference between these two scales (with and without the vertical scale items). This difference is used as the basis for relating grade-level Rasch scales to create the vertical scale.

To accomplish this, the mean/mean method was used to determine the equating constants and to place all the items administered to students within the same grade level (base-test, on-grade-level, and off-grade-level vertical scale items) on the same Rasch scale (Kolen & Brennan, 2004). The equating constants were then determined by finding the difference between the

means of the base-test items ($\bar{b}_{BTonly}$) from the Step 1 calibration and the mean of the base-test items ($\bar{b}_{BTandVS}$) from the Step 2 calibration.

$$C^* = \bar{b}_{BTonly} - \bar{b}_{BTandVS} \quad (2)$$

This difference (C^*) is the equating constant that is added to all the Rasch item difficulties for the vertical scale items from the second calibration step.

Stage 2 ("Across the Columns")

Once all the items within a grade level were on the same Rasch scale, adjacent grade vertical scale constants were calculated by comparing the mean Rasch item difficulties for the vertical scale items across adjacent grades (mean/mean method) using the vertical scale items. Vertical scale constants were computed in three ways:

1. using only lower-grade-level vertical scale items,
2. using only upper-grade-level vertical scale items, and
3. combining lower- and upper-grade-level vertical scale items.

Figure 8 illustrates the three ways to compute the vertical scale constants for adjacent grades using the mean/mean method.

1. To compute the adjacent grade vertical scale constants "across the columns" for grades 3 and 4, based on the *lower-grade-level* vertical scale items, the grade 3 vertical scale items are used. The mean Rasch item difficulty for the on-grade-level grade 3 vertical scale items in the grade 3 column is compared to the mean Rasch item difficulty for the lower-grade-level grade 3 vertical scale items in the grade 4 column.
2. To compute the adjacent grade vertical scale constants "across the columns" for grades 3 and 4, based on the *upper-grade-level* vertical scale items, the grade 4 vertical scale items are used. The mean Rasch item difficulty for the on-grade-level grade 4 vertical scale items in the grade 3 column is compared to the mean Rasch item difficulty for the upper-grade-level grade 4 vertical scale items in the grade 4 column.
3. To compute the adjacent grade vertical scale constants "across the columns" for grades 3 and 4, based on the *upper- and lower-grade-level* vertical scale items, all grade 3 and grade 4 vertical scale items are used. The mean Rasch item difficulty for the on-grade-level grade 3 vertical scale items along with the upper-grade-level grade 4 items in the grade 3 column is compared to the mean Rasch item difficulty for the lower-grade-level grade 3 vertical scale items and the on-grade-level grade 4 vertical scale items in the grade 4 column.

Estimating vertical scaling constants in three different ways provides options for creating the vertical scale and provides evidence for convergence or divergence of the vertical scaling results. Therefore, the best trajectory line between grade levels can be identified.

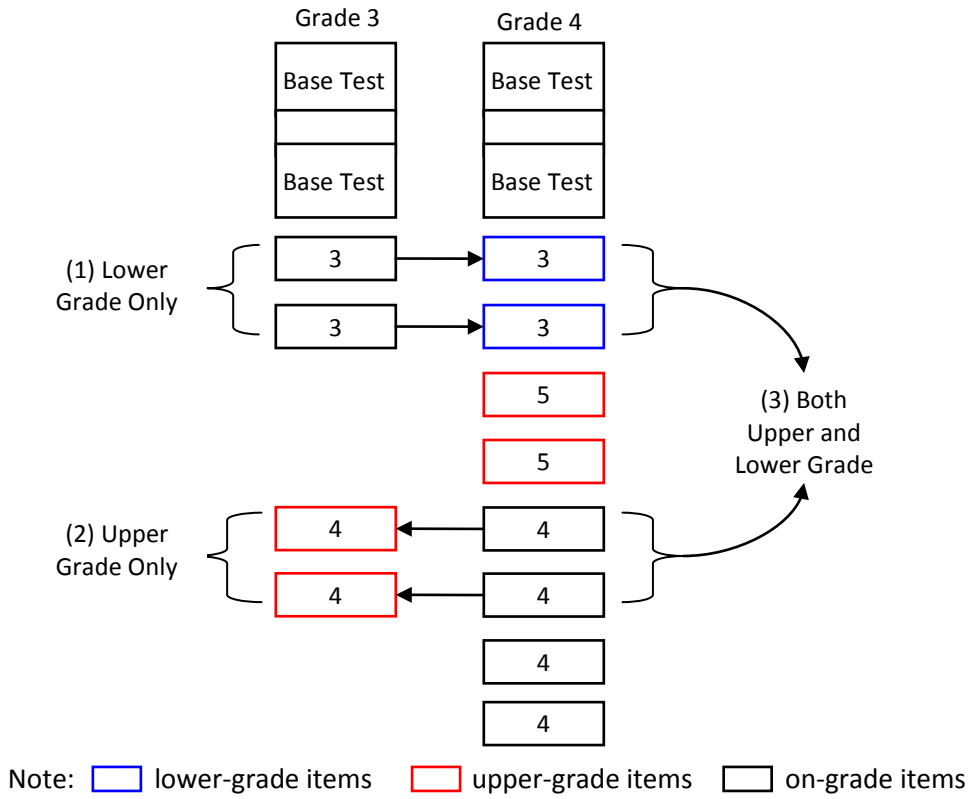


Figure 8. Illustration of Adjacent Grade Level Constants for STAAR English Grades 3 and 4 Reading "Across the Columns"

The formulas for computing the adjacent grade vertical scale constants for STAAR reading and mathematics are listed in Tables 10 and 11. The adjacent grade level vertical scale constants are determined by the difference between the means of the Rasch item difficulties of adjacent grades. For example, the adjacent grade vertical scale constant for grades 7 and 8, denoted VS_{78}, is defined as the difference between the mean Rasch item difficulty for the vertical scale items when calibrated and scaled with the grade 8 test, denoted $MeanVS_{GR8}$, and the mean Rasch item difficulty for the vertical scale items when calibrated and scaled with the grade 7 test, denoted $MeanVS_{GR7}$.

Table 10. Formulas for the STAAR 3–8 English Reading and Mathematics Adjacent Grade Vertical Scale Constants

Grade	Adjacent Grade Vertical Scale Constant (VS)
8	0
7	$VS_{78} = MeanVS_{GR8} - MeanVS_{GR7}$
6	$VS_{67} = MeanVS_{GR7} - MeanVS_{GR6}$
5	$VS_{56} = MeanVS_{GR6} - MeanVS_{GR5}$
4	$VS_{45} = MeanVS_{GR5} - MeanVS_{GR4}$
3	$VS_{34} = MeanVS_{GR4} - MeanVS_{GR3}$

Table 11. Formulas for the STAAR 3–5 Spanish Reading Adjacent Grade Vertical Scale Constants

Grade	Adjacent Grade Vertical Scale Constant (VS)
5	0
4	$VS_{45} = MeanVS_{GR5} - MeanVS_{GR4}$
3	$VS_{34} = MeanVS_{GR4} - MeanVS_{GR3}$

Stage 3 (Aggregating the Adjacent Grade Vertical Scale Constants)

After finding the vertical scale constants between adjacent grades, cumulative vertical scale constants were defined from the anchor grade to any grade levels that were not adjacent to the anchor grade. The anchor grade level was grade 8 for STAAR English reading and mathematics and grade 5 for STAAR Spanish reading. These grade levels were selected as the anchor grades to be consistent with the goals of the STAAR program of aligning the content standards, curriculum standards, and performance standards toward readiness for success in future grades.

Tables 12 and 13 list the formulas for computing the cumulative vertical scale constant (CVS) for STAAR reading and mathematics. The vertical scale constant at the anchor grades was set at zero, and the cumulative vertical scale constant at the other grades was calculated based on that end point. The cumulative vertical scale constants for grade 3–7 were based on the aggregate of the adjacent grade vertical scale constants. For example, the cumulative vertical scale constant for STAAR grade 5 (CVS_{58}) is the sum of the adjacent grade vertical scale constants for grades 5 and 6 (VS_{56}), grades 6 and 7 (VS_{67}), and grades 7 and 8 (VS_{78}).

Table 12. Formulas for the STAAR 3–8 English Reading and Mathematics Cumulative Vertical Scale Constants

Grade	Cumulative Vertical Scale Constant
8	$CVS = 0$
7	$CVS_{78} = 0+VS_{78}$
6	$CVS_{68} = 0+VS_{78}+VS_{67}$
5	$CVS_{58} = 0+VS_{78}+VS_{67}+VS_{56}$
4	$CVS_{48} = 0+VS_{78}+VS_{67}+VS_{56}+VS_{45}$
3	$CVS_{38} = 0+VS_{78}+VS_{67}+VS_{56}+VS_{45}+VS_{34}$

Table 13. Formulas for the STAAR 3–5 Spanish Reading Cumulative Vertical Scale Constants

Grade	Cumulative Vertical Scale Constant
5	$CVS = 0$
4	$CVS_{45} = 0 + VS_{45}$
3	$CVS_{35} = 0 + VS_{45} + VS_{34}$

Chapter 3: Results

This chapter provides a summary of the results for the vertical scale study and includes the following:

- Sample Size
- Evaluating the Vertical Scale Common Items
- Vertical Scale Common Items
- Vertical Scale Constants

Sample Size

Tables 14 through 16 list the number of students included in the vertical scale study overall and for each vertical scale form for STAAR mathematics, STAAR English reading, and STAAR Spanish reading, respectively. The test forms were spiraled at the student level in order to obtain a representative sample for each form. The number of students per vertical scale form was based on the overall student sample and the number of test forms. The number of test forms varies for grades and subjects due to some tests needing more forms to field test more items to support the primary administration and multiple retest administrations. Specifically, STAAR English grade 5 reading and mathematics, STAAR Spanish grade 5 reading, and STAAR grade 8 reading and mathematics tests had fewer students per form because there were more test forms for these grades than for grades 3, 4, 6, or 7. The number of students per vertical scale form for STAAR 3–8 mathematics ranged from 5,261 to 8,854. The number of students per vertical scale form for STAAR English 3–8 reading ranged from 5,324 to 8,635. The number of students per vertical scale form for STAAR Spanish 3–5 reading ranged from 350 to 1,182. The student population for STAAR Spanish reading decreased as grade level increased because students transitioned to the English version of the reading test as their language proficiency improves.

Table 14. Sample Size for STAAR Mathematics Overall and by Vertical Scale Form

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	Overall	337,030	346,183	337,803	344,858	322,888	312,088
Vertical Scale Form	1	8,826	8,018	5,287	8,000	7,373	5,277
	2	8,798	8,038	5,300	8,023	7,360	5,304
	3	8,854	8,059	5,265	8,030	7,402	5,302
	4	8,799	8,055	5,322	7,998	7,365	5,273
	5	--	8,023	5,292	8,016	7,412	--
	6	--	8,022	5,261	8,018	7,364	--
	7	--	8,005	5,308	8,005	7,407	--
	8	--	7,964	5,328	7,935	7,333	--

Table 15. Sample Size for STAAR English Reading Overall and by Vertical Scale Form

		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	Overall	327,719	334,418	348,762	354,316	347,825	310,278
Vertical Scale Form	1	8,610	7,775	5,489	8,218	7,973	5,343
	2	8,574	7,791	5,473	8,259	7,971	5,358
	3	8,635	7,786	5,471	8,253	7,989	5,332
	4	8,598	7,805	5,518	8,237	7,987	5,324
	5	--	7,800	5,480	8,216	8,040	--
	6	--	7,752	5,461	8,250	8,005	--
	7	--	7,772	5,476	8,243	8,036	--
	8	--	7,718	5,496	8,176	7,955	--

Table 16. Sample Size for STAAR Spanish Reading Overall and by Vertical Scale Form

		Grade 3	Grade 4	Grade 5
	Overall	36,199	23,228	9,569
Vertical Scale Form	1	1,182	871	355
	2	1,159	854	350
	3	1,139	871	355
	4	1,157	854	374
	5	--	862	--
	6	--	847	--
	7	--	836	--
	8	--	814	--

Evaluating the Vertical Scale Common Items

Vertical scaling assumes that while items may be of greater or lesser difficulty across grade levels, their relative difficulty to each other will generally be consistent. For example, it would make sense instructionally for long division to be more difficult for 4th graders than for 5th graders; however, it would not make sense instructionally for addition to be more difficult than long division at any grade level. If items rank order very differently in terms of their difficulty at different grade levels or if their statistical properties (such as fit to the Rasch model) appear unexpectedly inconsistent across grade levels, it may be desirable to drop certain items from the vertical scale common-item set prior to determining the final vertical scale constants between grade-levels.

The goal is to have the most stable and accurate vertical scaling constants possible. Therefore, the evaluation of the vertical scale common items needs to balance elimination of vertical scale items based on deviation from model fit and perfect rank ordering, against having fewer items

with which to calculate the vertical scale constant and maintain appropriate content representation. Therefore, vertical scale items should be dropped when the gain in stability of the vertical scaling constants is less than the estimation bias introduced from the inclusion of poor fitting items.

Some of the item statistics resulting from the calibrations can be used as an indicator of model misfit. Items that show model misfit should be identified for further evaluation. The vertical scale study used the Rasch mean-square infit statistic which is sensitive to items with unexpected response distributions when evaluated using students with Rasch-based performance estimates (θ) similar to the item's Rasch item difficulty (Linacre, 2001). The Rasch mean-square infit statistic is a chi-square statistic with 1.0 as the expected value. This statistic is provided by the calibration software when estimating the Rasch item difficulties. If the Rasch mean-square infit is unexpectedly large or small, then the item may be eliminated from the vertical scale item set. Using a conservative criterion, if a vertical scale item's Rasch mean-square infit statistic is between 0.80 and 1.20, the item fits the model well and it can be included in the final set of vertical scale items. Since each vertical scale item is calibrated with two grade levels, two Rasch mean-square infit statistics were computed.

There were six vertical scale items dropped from the vertical scale common-item sets due to low or high mean square infit statistics. Of the misfitting items, one grade 8 reading item had infit statistics above 1.20 when calibrated on-grade-level with the grade 8 student data and off-grade-level with the grade 7 student data. The remaining items had infit statistics outside the acceptable range when calibrated with off-grade-level student data. Table 17 lists the number of vertical scale items per grade and subject that were removed from the common-item sets due to poor item-model fit. The grade associated with the item is listed in parenthesis next to the count of the number of items dropped. For example, for STAAR English reading a grade 4 item had a Rasch mean-square infit statistic outside the acceptable range when calibrated with the grade 5 (GR 5) student data.

Table 17. Vertical Scale Items Dropped Due to Model Fit

	Grade					
Subject	3	4	5	6	7	8
Mathematics	0	0	0	1 (GR 5)	0	0
English Reading	1 (GR 4)	1 (GR 5)	0	0	1 (GR 8)*	1 (GR 8)*
Spanish Reading	0	2 (GR 5)	0	-	-	-

Note: Items removed where Rasch mean-square infit < 0.80 or mean-square infit > 1.20

*Represents the same item being identified as misfitting for both populations.

Other criteria used to evaluate vertical scale items include comparisons of Rasch item difficulties for the same item, when taken by students as on-grade and off-grade items. Scatterplots for the adjacent grade Rasch item difficulties were used to identify, evaluate, and possibly eliminate vertical scale items from the vertical scale common-item sets prior to scaling

the adjacent grade. The scatterplots were fitted with simple linear regression. The vertical scale items in common between adjacent grade levels were evaluated so that the lower grade level Rasch item difficulties for the vertical scale common items were regressed on the upper grade level Rasch item difficulties for the vertical scale common items. Figure 9 is an example scatterplot for the vertical scale common-item set between grades 6 and 7 mathematics. The horizontal axis represents the Rasch item difficulties (BASERID06) when taken by students in grade 6 mathematics and the vertical axis represents the Rasch item difficulties (BASERID07) for the same items when taken by students in grade 7 mathematics.

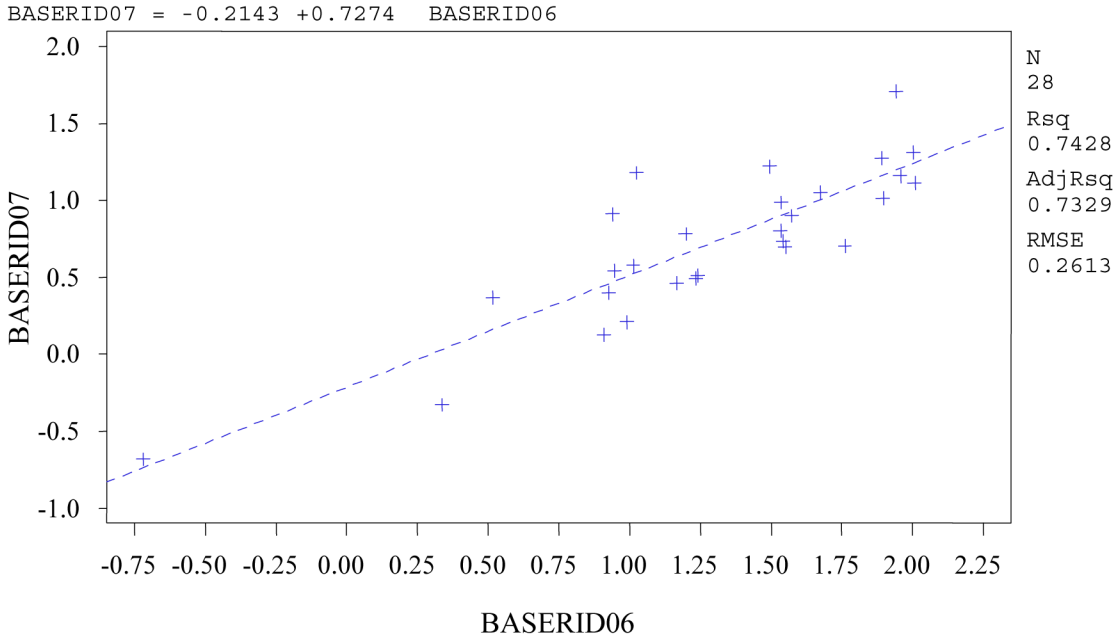


Figure 9. Scatterplot of Adjacent Grade Level Rasch Item Difficulties for STAAR Mathematics Grades 6 and 7

The residuals (the observed value minus the predicted value based on the regression equation) are used to evaluate whether a common item is not fitting the regression model. The residuals are denoted $\hat{\varepsilon}$. The statistic used for this evaluation is the externally studentized residual, which is reported in estimated standard deviation units, σ, based on the standard error of the predicted values, without including the item of interest in the estimate of the standard error. Removing the item from the estimate of the standard error results in an externally studentized residual. Otherwise, inclusion of the item results in an internally studentized residual.

Since the actual number of items eliminated does have an impact on the ability to create the vertical scale, it is also vital that the final vertical scale common-item set covers the content as was originally intended. Tables 20 through 22 display, by reporting category, the percentage of items for each adjacent grade set that were in the initial vertical scale common-item set, and the percentage in the final vertical scale common-item set after the removal of misfitting items and outliers. The content representation is shown for the adjacent grade levels combining the vertical scale common items. For STAAR 3–8 mathematics, the percentages are fairly similar between the initial and final vertical scale common-item sets.

Table 20. STAAR 3–8 Mathematics Adjacent Grade Level Vertical Scale Items Content Representation

	Adjacent Grade Levels									
Reporting	3–4		4–5		5–6		6–7		7–8	
Category	Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final
1	32%	33%	36%	33%	33%	35%	32%	35%	21%	24%
2	18%	19%	14%	15%	15%	12%	25%	23%	29%	32%
3	21%	22%	18%	19%	15%	15%	14%	15%	21%	24%
4	14%	15%	18%	19%	19%	19%	14%	12%	14%	8%
5	14%	11%	14%	15%	19%	19%	14%	15%	14%	12%
Total Items	28	27	28	27	27	26	28	26	28	25

For STAAR 3–8 English reading, the percentages are fairly similar between the initial and final vertical scale common-item sets. The lack of items for reporting category 2 for the adjacent grades 4–5 and the lack of reporting category 3 for the adjacent grades 5–6 are due to the size of the STAAR 3–8 reading item banks and the dependency on passage-based items. The STAAR passages are developed to assess one to two reporting categories; therefore, a passage shared between the adjacent grades may result in zero items for one of the reporting categories. The item bank and the psychometric requirements for items resulted in the lack of items for some of the adjacent grade level vertical scale common-item sets.

Table 21. STAAR 3–8 Reading Adjacent Grade Level Vertical Scale Items Content Representation

	Adjacent Grade Level									
Reporting	3–4		4–5		5–6		6–7		7–8	
Category	Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final
1	19%	21%	14%	12%	15%	12%	19%	20%	16%	17%
2	63%	62%	0%	0%	85%	88%	38%	37%	44%	45%
3	19%	17%	86%	88%	0%	0%	44%	43%	41%	38%
Total Items	32	29	28	26	27	26	32	30	32	29

For STAAR 3–5 Spanish reading, the percentages are fairly similar between the initial and final vertical scale common-item sets.

Table 22. STAAR Spanish 3–5 Reading Adjacent Grade Level Vertical Scale Items Content Representation

Reporting Category	Adjacent Grade Level			
	3–4		4–5	
	Initial	Final	Initial	Final
1	23%	21%	21%	22%
2	58%	62%	21%	22%
3	19%	17%	57%	57%
Total Items	31	29	28	23

Final Vertical Scale Common-Item Set

The vertical scale common-item sets were the items in common between adjacent grade levels after dropping items due to model misfit and studentized residuals. The vertical scale common items between adjacent grades were used to compute three vertical scale constants based on using only lower-grade level items, only upper-grade level items, and both lower- and upper-grade level items. The vertical scale constants were computed using the formulas in Tables 10 and 11. Tables 23 through 25 list the three vertical scale constants for the adjacent grade levels for STAAR grades 3–8 mathematics, STAAR English grades 3–8 reading, and STAAR Spanish grades 3–5 reading, respectively, for the vertical scale common-item sets.

Table 23. STAAR 3–8 Mathematics Adjacent Grade Vertical Scale Constants

Adjacent Grade Levels	Lower Grade Items		Upper Grade Items		All Items	
	N	Constant	N	Constant	N	Constant
3–4	13	0.6658	14	0.7058	27	0.6865
4–5	13	0.3986	14	0.5024	27	0.4524
5–6	13	0.7652	13	0.7100	26	0.7376
6–7	12	0.4637	14	0.4831	26	0.4742
7–8	12	0.4533	13	0.4254	25	0.4388

Table 24. STAAR English 3–8 Reading Adjacent Grade Vertical Scale Constants

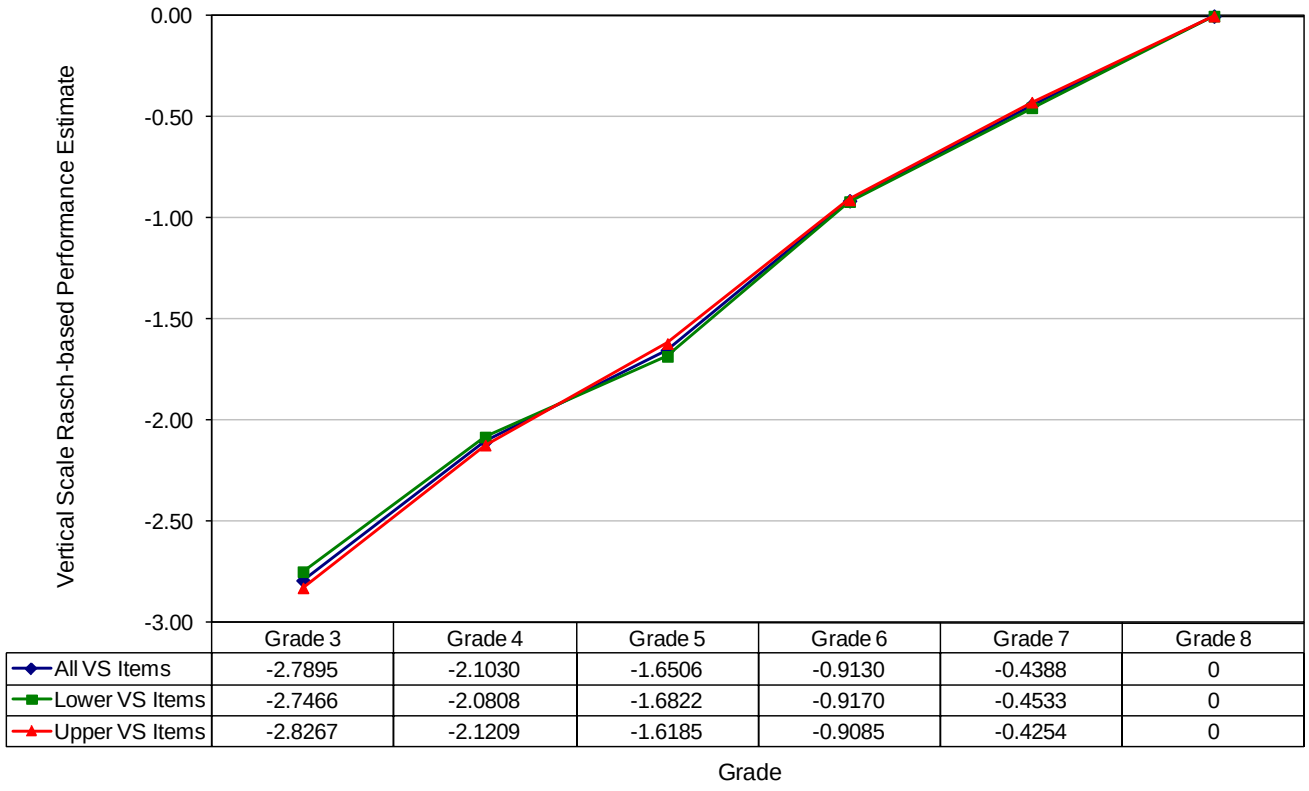
Adjacent Grade Levels	Lower Grade Items		Upper Grade Items		All Items	
	N	Constant	N	Constant	N	Constant
3–4	14	0.6163	15	0.6241	29	0.6203
4–5	12	0.2998	14	0.3499	26	0.3268
5–6	12	0.4268	14	0.3597	26	0.3907
6–7	15	0.4550	15	0.4606	30	0.4578
7–8	16	0.2126	13	0.2070	29	0.2101

Table 25. STAAR Spanish 3–5 Reading Adjacent Grade Vertical Scale Constants

Adjacent Grade Levels	Lower Grade Items		Upper Grade Items		All Items	
	N	Constant	N	Constant	N	Constant
3–4	15	0.6273	14	0.5320	29	0.5813
4–5	10	0.2709	13	0.2694	23	0.2700

The three methods of computing the adjacent grade vertical scale constants were evaluated by comparing the cumulative vertical scale constants by setting the anchor grade to zero and aggregating the vertical scale constants from the upper- to the lower-grade levels. The cumulative vertical scale constants were computed using the formulas in Tables 12 and 13. Figures 10 through 12 show the relationship between lower, upper, and combined vertical scale common-item sets by listing the cumulative vertical scale constants for each method and plotting the constants by grade level.

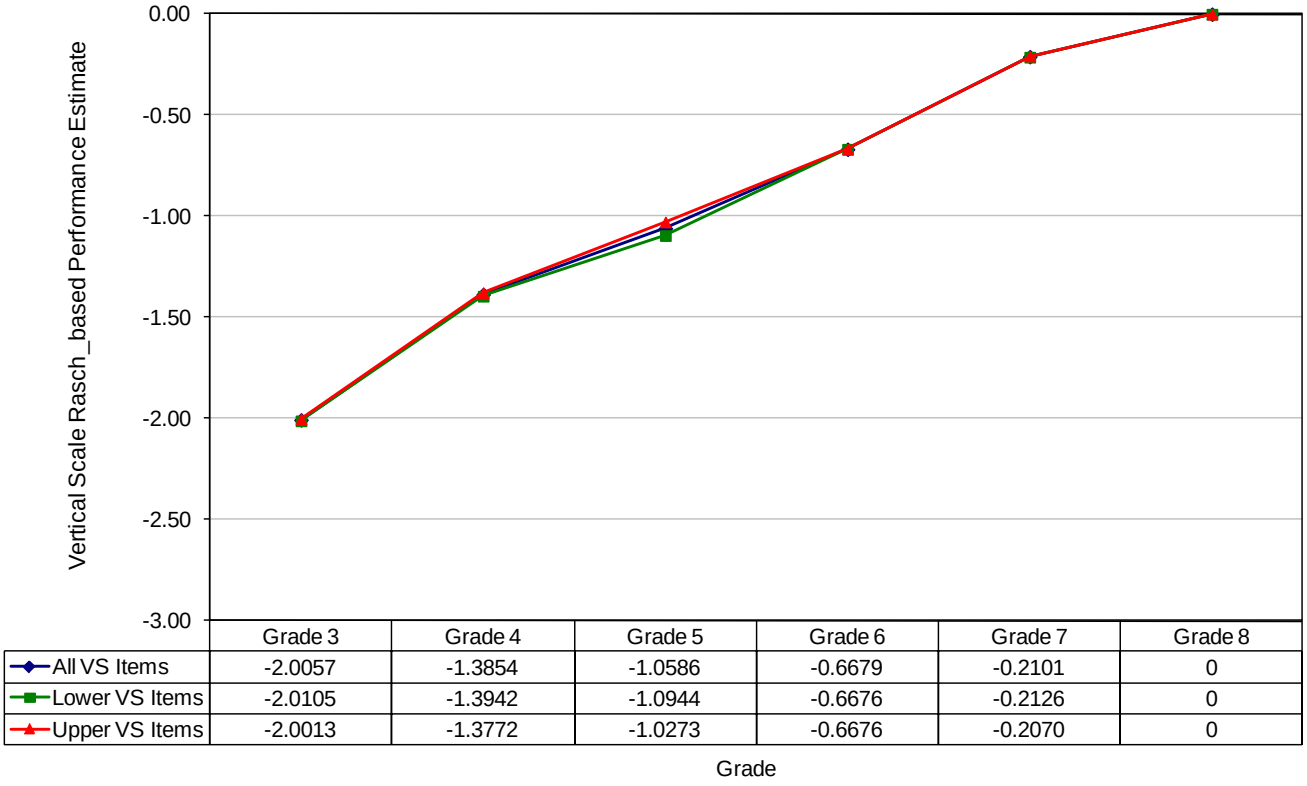
Figure 10 shows the STAAR 3–8 mathematics cumulative vertical scale constants for each method. The trend across the three methods is similar. The vertical scale increases from grade 3 to grade 8 indicating increasing difficulty of the assessments. The cumulative vertical scale constants between the three methods are similar at each grade level with differences ranging from 0.004 to 0.0801. The cumulative vertical scale constants have the largest difference between the upper-level items and the lower-level items at grade 5 (0.0637) and grade 3 (0.0801). The cumulative vertical scale combined constants (lower- and upper-grade-level items) fall between the lower-grade-level constant and the upper-grade-level constant at each grade level. The lines are very similar except for the slight departure at grades 3 and 5.



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All VS Items	-2.7895	-2.1030	-1.6506	-0.9130	-0.4388	0
Lower VS Items	-2.7466	-2.0808	-1.6822	-0.9170	-0.4533	0
Upper VS Items	-2.8267	-2.1209	-1.6185	-0.9085	-0.4254	0

Figure 10. STAAR 3–8 Mathematics Cumulative Vertical Scale Constants for Lower, Upper, and Combined Common-Item Sets

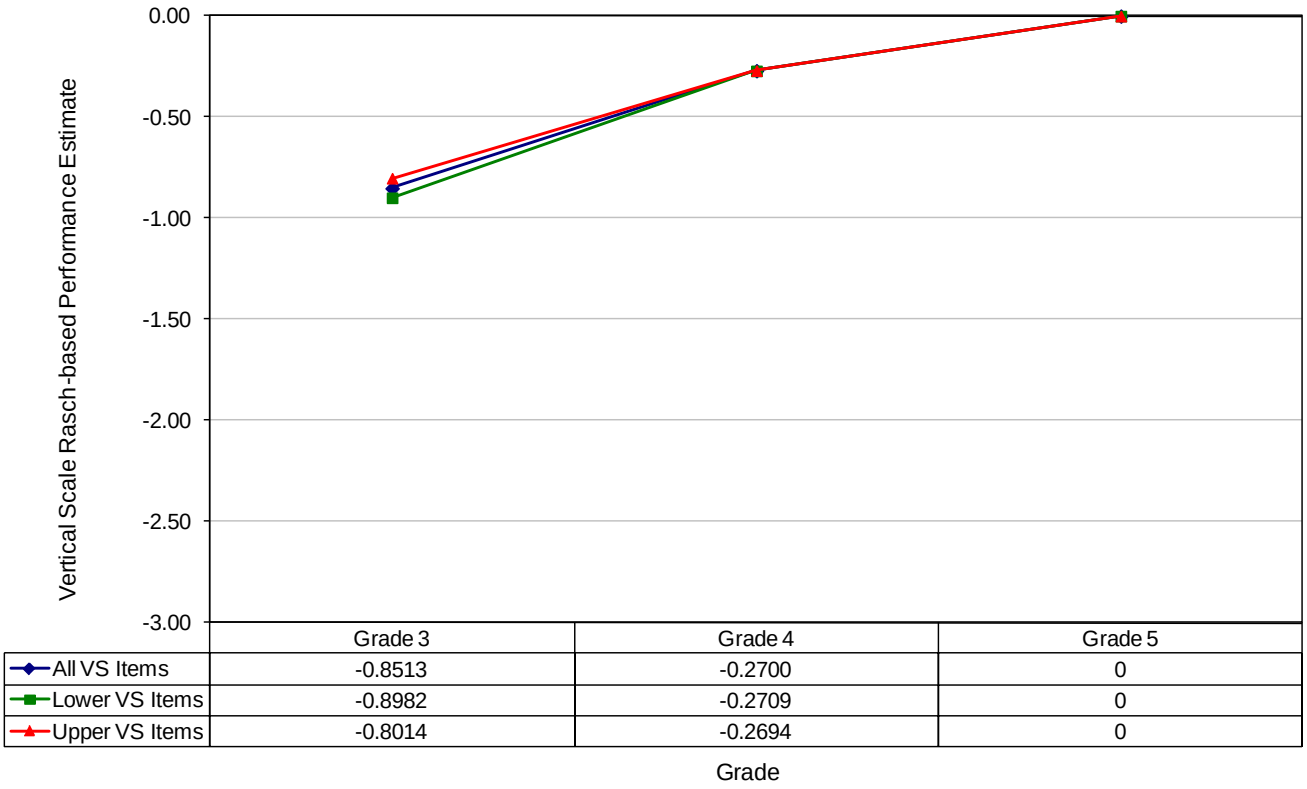
Figure 11 shows the STAAR 3–8 English reading cumulative vertical scale constants for each method. The trend across the three methods is similar. The vertical scale increases from grade 3 to grade 8 indicating the increasing difficulty of the assessments. The cumulative vertical scale constants between the three methods are similar at each grade level with differences ranging from 0.0 to 0.0671. The cumulative vertical scale constants have the largest difference between the upper-level items and the lower-level items at grade 5 (0.0671). The cumulative vertical scale combined constants (lower- and upper-grade-level items) fall between the lower-grade-level constant and the upper-grade-level constant at each grade level. The lines are very similar except for the slight departure at grade 5.



	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All VS Items	-2.0057	-1.3854	-1.0586	-0.6679	-0.2101	0
Lower VS Items	-2.0105	-1.3942	-1.0944	-0.6676	-0.2126	0
Upper VS Items	-2.0013	-1.3772	-1.0273	-0.6676	-0.2070	0

Figure 11. STAAR 3–8 English Reading Vertical Scale Constants for Lower, Upper, and Combined Common-Item Sets

Figure 12 shows the STAAR Spanish 3–5 reading cumulative vertical scale constants for each method. The trend across the three methods is similar. The vertical scale increases from grade 3 to grade 5 indicating the increasing difficulty of the assessments. The cumulative vertical scale constants between the three methods are similar at each grade level with differences ranging from 0.0 to 0.0968. The cumulative vertical scale constants have the largest difference between the upper-level items and the lower-level items at grade 3 (0.0968). The cumulative vertical scale combined constants (lower- and upper-grade-level items) fall between the lower-grade-level constant and the upper-grade-level constant at each grade level. STAAR Spanish 3–5 reading appears to show slight differences in the trajectories between grades 3 and 4.



	Grade 3	Grade 4	Grade 5
All VS Items	-0.8513	-0.2700	0
Lower VS Items	-0.8982	-0.2709	0
Upper VS Items	-0.8014	-0.2694	0

Figure 12. STAAR Spanish 3–5 Reading Vertical Scale Constants for Lower, Upper, and Combined Common-Item Sets

Based on a review of Figures 10 through 12, a decision was made to use both the lower-grade-level and upper-grade-level items in the vertical scale common-item set to define the STAAR vertical scales. The trends in the three methods were consistent. Using all the vertical scale items resulted in a larger number of common items and better content representation.

Cumulative Vertical Scaling Constant

As indicated earlier, the mean/mean equating procedure was used to find the vertical linking constants between adjacent grade levels. After finding the linking constant between adjacent grades, a cumulative linking constant was defined from the anchor grade to all other grade levels. For example, at grade 6, the *vertical scale constant* (between grades 6 and 7) would be the difference between the mean vertical scale item difficulties for grade 6 and grade 7. On the other hand, the *cumulative vertical scale constant* between grade 6 and the anchor grade level (grade 8) is the vertical scale constant between grades 6 and 7 *plus* the vertical scale constant between grades 7 and 8 (see Table 26). The same final vertical scale constants will be used for future STAAR administrations.

Table 26. Final Cumulative Vertical Scale Constants

Grade	STAAR 3–8 Mathematics	STAAR 3–8 English Reading	STAAR 3–5 Spanish Reading
	Cumulative VS Constant	Cumulative VS Constant	Cumulative VS Constant
8	0	0	-
7	-0.4388	-0.2101	-
6	-0.9130	-0.6679	-
5	-1.6506	-1.0586	0
4	-2.1030	-1.3854	-0.2700
3	-2.7895	-2.0057	-0.8513

Note. Cumulative constants based on *all* vertical scale items (upper and lower grade level).

Chapter 4: Evaluation

The evaluation of a vertical scale is not straight forward. There are no firm guidelines by which one can judge all such scales. However, for the STAAR vertical scales, several criteria have been identified for evaluating the scale. The development of these criteria was in part based on consultation with Dr. Michael Kolen, a member of the Texas Technical Advisory Committee, and from a paper by Patz (2007). This chapter summarizes the criteria for evaluating the vertical scale results and includes the following:

- Progression in Difficulty Across Grades
- Vertical Scale Means and Standard Deviations
- Relationship between Vertical Scale Item Sets

Progression in Difficulty Across Grades

It seems a reasonable assumption that, when comparing tests designed to assess similar content at different grade levels, the difficulty of the upper grade test would be higher than the difficulty of the lower grade test. Therefore, an initial check of the STAAR vertical scales is made to verify that this progression exists.

Figures 10 through 12 indicate that the vertical scales for STAAR English 3–8 mathematics and reading and STAAR Spanish 3–5 reading show upward trends, indicating that the average difficulty of the upper grade test is higher than the previous grade. For example, the difficulty of the grade 4 test is higher than the difficulty of the grade 3 test. Based on review of the "all" items lines in Figures 10 through 12, it appears that the STAAR vertical scales meet this criterion.

Another reasonableness check on a vertical scale is the performance of the vertical scale items. At the item level, it seems reasonable that an item should not perform very differently, relative to the other items, across grades. One way to look at across-grade differences is to examine the correlation coefficient between Rasch item difficulties for vertical scale items in adjacent grade levels. As Patz (2007) noted, "high degrees of correlation suggest that the examinees and/or items would be ordered the same way on adjacent test levels, which may be taken as a degree of validation that the vertical scale is appropriate." (p.18)

Table 27 provides the correlation between the Rasch item difficulties for the final adjacent grade vertical scale common-item sets. The correlations were high and positive, with the lowest being 0.90 in the adjacent grade levels for 4–5 and 6–7 mathematics and the highest being 0.98 for the adjacent grade levels for 4–5, 5–6, 6–7, and 7–8 reading.

Table 27. Correlation for Vertical Scale Items in Common-Item Sets for Adjacent Grades

	Adjacent Grade Levels				
Subject	3–4	4–5	5–6	6–7	7–8
Mathematics	0.91	0.90	0.93	0.90	0.96
English Reading	0.94	0.98	0.98	0.98	0.98
Spanish Reading	0.96	0.96	-	-	-

Vertical Scale Means and Standard Deviations

Vertical scale Rasch-based performance estimate (θ) means should increase across grade levels in a regular pattern. Just as the vertical scale constants should progress in difficulty, it is reasonable to assume that the application of these constants should affect the population of interest in a similar manner. Vertical scale Rasch-based performance estimate (θ) standard deviations should not have large differences or systematic increases/decreases from grade to grade. The standard deviation reflects the variability in the student population and is expected to be similar within a content area across grades. Deviations from the expected trends in the means or differences in the standard deviation of the vertical scale Rasch-based performance estimate (θ) require additional evaluation.

Tables 28 through 30 provide summary statistics for the vertical scale Rasch-based performance estimate (θ) based on the student populations tested in 2012 for STAAR English 3–8 mathematics and reading and STAAR Spanish 3–5 reading, respectively. In addition Figures 13 through 15 illustrate the means and standard deviations of the vertical scale Rasch-based performance estimates (θ) across grade levels. As expected, the means of the vertically scaled Rasch-based performance estimates (θ) increased across grade levels.

For STAAR 3–8 mathematics, the standard deviation of the vertical scale Rasch-based performance estimate (θ) for grade 8 mathematics is smaller in relation to the other grade levels. Further examination of the student population revealed that the change in the STAAR administrations resulted in students taking the end-of-course Algebra I test rather than the grade 8 mathematics test if they were enrolled in Algebra I in grade 8. This resulted in a slightly smaller standard deviation. The same trend in the standard deviation is not observed for grade 8 reading.

Table 28. Summary Statistics for STAAR 3–8 Mathematics Vertical Scale Rasch-based Performance Estimate (θ) for 2012

Grade	N	Mean	Standard Deviation	Minimum	Maximum	Range
3	337,030	-0.70	1.24	-6.89	3.67	10.56
4	346,183	-0.02	1.16	-6.28	4.33	10.61
5	337,804	0.45	1.24	-5.83	4.74	10.57
6	344,859	0.77	1.34	-5.23	5.48	10.71
7	322,889	0.84	1.14	-4.95	5.85	10.81
8	312,088	1.17	1.05	-4.50	6.28	10.78

Table 29. Summary Statistics for STAAR English 3–8 Reading Vertical Scale Rasch-based Performance Estimate (θ) for 2012

Grade	N	Mean	Standard Deviation	Minimum	Maximum	Range
3	327,719	-0.86	1.21	-6.92	3.41	10.33
4	334,418	-0.06	1.12	-6.20	4.21	10.41
5	313,529	0.20	1.09	-6.11	4.50	10.61
6	354,317	0.59	1.10	-5.75	4.89	10.64
7	347,826	1.02	1.04	-5.11	5.38	10.49
8	310,279	1.32	1.10	-5.03	5.70	10.73

Table 30. Summary Statistics for STAAR Spanish 3–5 Reading Vertical Scale Rasch-based Performance Estimate (θ) for 2012

Grade	N	Mean	Standard Deviation	Minimum	Maximum	Range
3	36,200	-0.25	1.03	-5.64	4.50	10.14
4	23,230	0.36	1.04	-5.23	5.13	10.35
5	9,573	0.87	1.00	-4.85	5.61	10.46

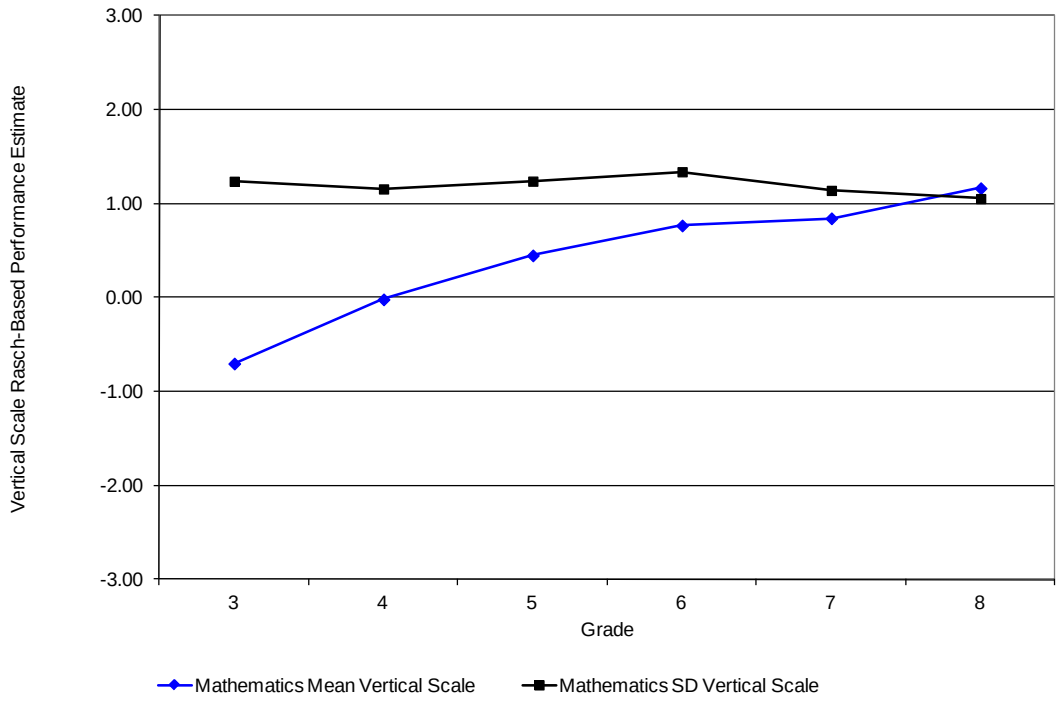


Figure 13. STAAR 3–8 Mathematics Vertical Scale Rasch-based Performance Estimate (θ) Mean and Standard Deviation for 2012

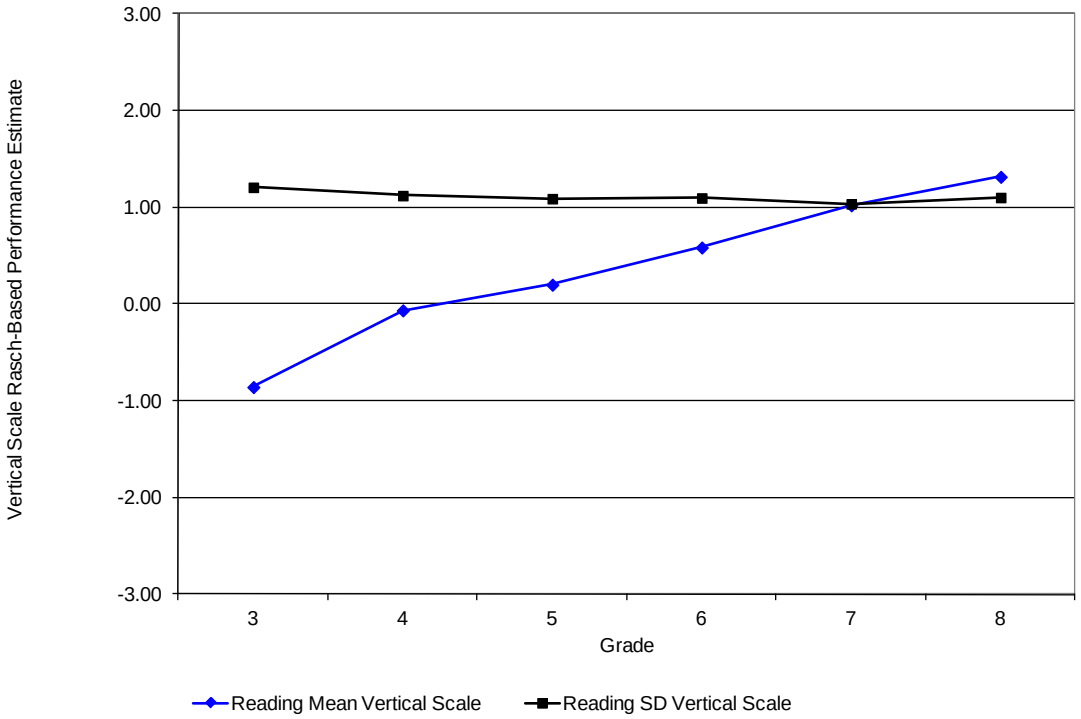


Figure 14. STAAR English 3–8 Reading Vertical Scale Rasch-based Performance Estimate (θ) Mean and Standard Deviation for 2012

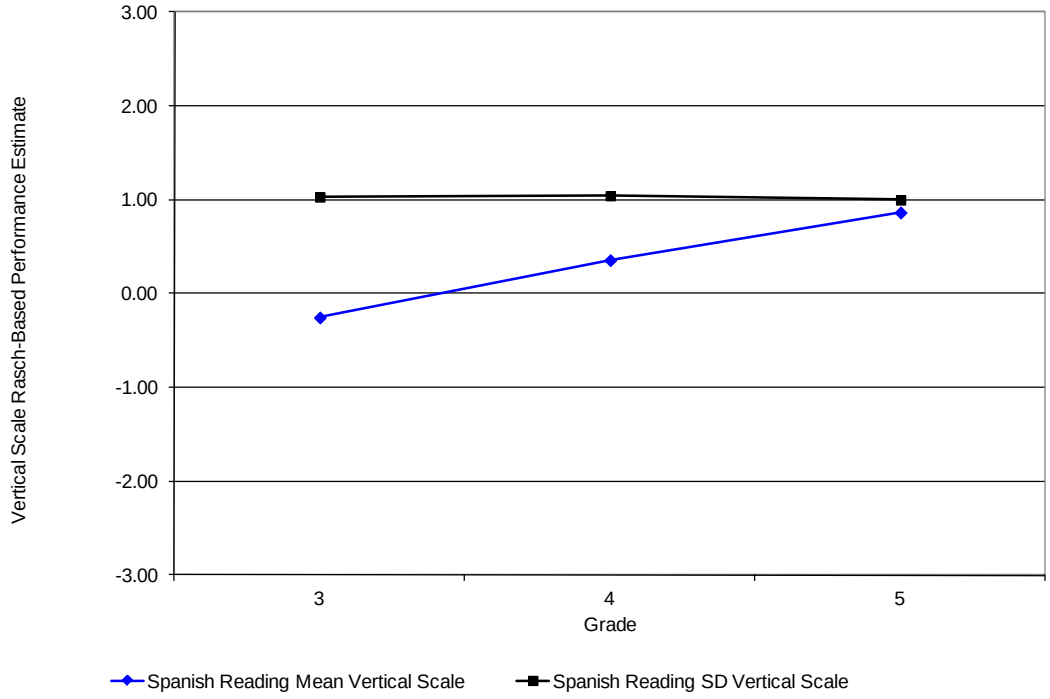


Figure 15. STAAR Spanish 3–5 Reading Vertical Scale Rasch-based Performance Estimate (θ) Mean and Standard Deviation for 2012

Relationship between Vertical Scale Item Sets

The relationship between lower, upper, and combined vertical scale item sets should be regular (i.e., vertical scale line plots from each group are increasing) and the differences between the vertical scales derived from each vertical scale item set should be minimal. Estimating vertical scale constants in three different ways provides options for creating the vertical scale. This also provides a good way to cross-validate the vertical scale methodology. If the different vertical scale common-item sets provide a similar estimate of growth then the vertical scale constants should be very similar. If they provide different estimates of growth, then one would expect that differences would be approximately a constant due to the inclusion of the vertical scale items in establishing the mean/mean vertical scale constant. Based on the plots shown in Figures 10 through 12 it appears that the three different vertical scale common-item sets provide very similar results for the STAAR vertical scale.

The STAAR vertical scale results were reviewed by a Texas Technical Advisory Committee member, Michael Kolen, in August 2012. The vertical scale items dropped from the common-item sets were discussed in terms of the number of items for the vertical linking constant and stability of the Rasch item difficulties. The STAAR vertical scale results were deemed reasonable.

Chapter 5: Implementation

This section discusses the implementation of the vertical scale study for establishing the STAAR 3–8 scale score system and includes the following:

- Performance Categories
- Vertical Scaling Constants

Performance Categories

As part of the implementation of the general STAAR 3–8 assessment program, performance standards (or cut scores) were set for each assessment to establish three performance categories. For the general STAAR assessments, the labels for the performance categories are:

- *Level III: Advanced Academic Performance*
- *Level II: Satisfactory Academic Performance*
- *Level I: Unsatisfactory Academic Performance*

The policy definitions for each of the performance categories are as follows:

Level III: Advanced Academic Performance
Performance in this category indicates that students are well prepared for the next grade or course. They demonstrate the ability to think critically and apply the assessed knowledge and skills in varied contexts, both familiar and unfamiliar. Students in this category have a high likelihood of success in the next grade or course with little or no academic intervention.

Level II: Satisfactory Academic Performance
Performance in this category indicates that students are sufficiently prepared for the next grade or course. They generally demonstrate the ability to think critically and apply the assessed knowledge and skills in familiar contexts. Students in this category have a reasonable likelihood of success in the next grade or course but may need short-term, targeted academic intervention.

Level I: Unsatisfactory Academic Performance
Performance in this category indicates that students are inadequately prepared for the next grade or course. They do not demonstrate a sufficient understanding of the assessed knowledge and skills. Students in this category are unlikely to succeed in the next grade or course without significant, ongoing academic intervention.

Vertical Scaling Constants

The STAAR vertical scale results were implemented with the new scale score system for reporting students' scores. The vertically scaled STAAR scale scores represent linear

transformations of Rasch-based performance estimates (θ). Vertically scaled scores include three scaling constants: the slope (A), intercept (B), and the cumulative vertical scale constant (V_g). The cumulative vertical scale constant varies across each grade (g). The vertical scale scores (SC_θ) are computed using the following equation:

$$SC_\theta = A \times (\theta - V_g) + B \tag{4}$$

The scale score at the Level II cut is fixed for only the anchor grade (STAAR English grade 8 mathematics and reading or STAAR Spanish grade 5 reading) and the standard deviation is taken across all of the assessments. The A scaling constant is calculated as follows:

$$A = \frac{\sigma_{sc}}{\sigma_\theta} \tag{5}$$

In Equation (5), σ_{sc} represents the desired standard deviation of the scale across all assessments, while σ_θ represents the standard deviation of Rasch-based θ values among a sample group. For the STAAR 3–8 vertical scales, the sample group consisted of all students who took the assessment across the vertical scale in spring 2012.

The B scaling constant is calculated as follows:

$$B = SC_{Level\_II} - \frac{\sigma_{sc}}{\sigma_\theta} \times \theta_{Level\_II} \tag{6}$$

In Equation (6), SC_{Level_II} represents the desired scale score at the Level II cut for the final assessment in the vertical scale, and θ_{Level_II} represents the approved Level II performance standard (in Rasch units) for the final assessment in the vertical scale. As in Equation (5), σ_{sc} represents the desired standard deviation of the scale, while σ_θ represents the standard deviation of Rasch-based θ values in the sample group. Using Equation (4) and substituting Equation (5) for A and Equation (6) for B, the full STAAR vertical scaling equation is shown below.

$$SC_\theta = \frac{\sigma_{sc}}{\sigma_\theta} \times (\theta - V_g) + \left[SC_{Level\_II} - \frac{\sigma_{sc}}{\sigma_\theta} \times \theta_{Level\_II} \right] \tag{7}$$

For the STAAR English 3–8 mathematics and reading vertical scales, a scale score of 1700 represents the recommended Level II performance standard for the grade 8 assessment. In addition, those scales' standard deviations were set to 150. These values can be substituted into Equation (7) to provide a scaling equation specific to the mathematics and English reading vertical scaled assessments.

$$SC_\theta = \frac{150}{\sigma_\theta} \times (\theta - V_g) + \left[1700 - \frac{150}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (8)$$

For the STAAR Spanish grade 5 reading assessment, a scale score of 1582 represents the recommended Level II performance standard. This scale score is set to the equivalent value as the Level II performance standard on the STAAR English grade 5 reading assessment. The Spanish reading vertical scale's standard deviations was also set to 150. These values can be substituted into Equation (7) to provide a scaling equation specific to STAAR Spanish grades 3–5 reading vertical scale.

$$SC_\theta = \frac{150}{\sigma_\theta} \times (\theta - V_g) + \left[1582 - \frac{150}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (9)$$

It is important to note that although the Level II scale score cut is fixed for the highest grade in the vertical scale, the Level II cuts for the other assessments in the vertical scale will vary across grades. These Level II cuts, as well as the Level III cuts, do not vary over time. The fixed scale scores to be associated with the lower grades' Level II cuts (both phase-in and recommended) and all Level III cuts were calculated by substituting Level II and Level III-specific θ values into Equations (8) and (9) for each grade.

Figures 16 through 18 illustrate the Level II and Level III cut scores for mathematics, English reading, and Spanish reading, respectively. The STAAR vertical scales have the following characteristics:

- They range from approximately 600 to 2300 scale score points.
- The Level II cut score is 1700 for STAAR English grade 8 mathematics and reading.
- The Level II cut score is 1582 for STAAR Spanish grade 5 reading, which is the same for STAAR English grade 5 reading.
- Level II cut scores increase across grades within a content area.
- Level III cut scores increase across grades within a content area.

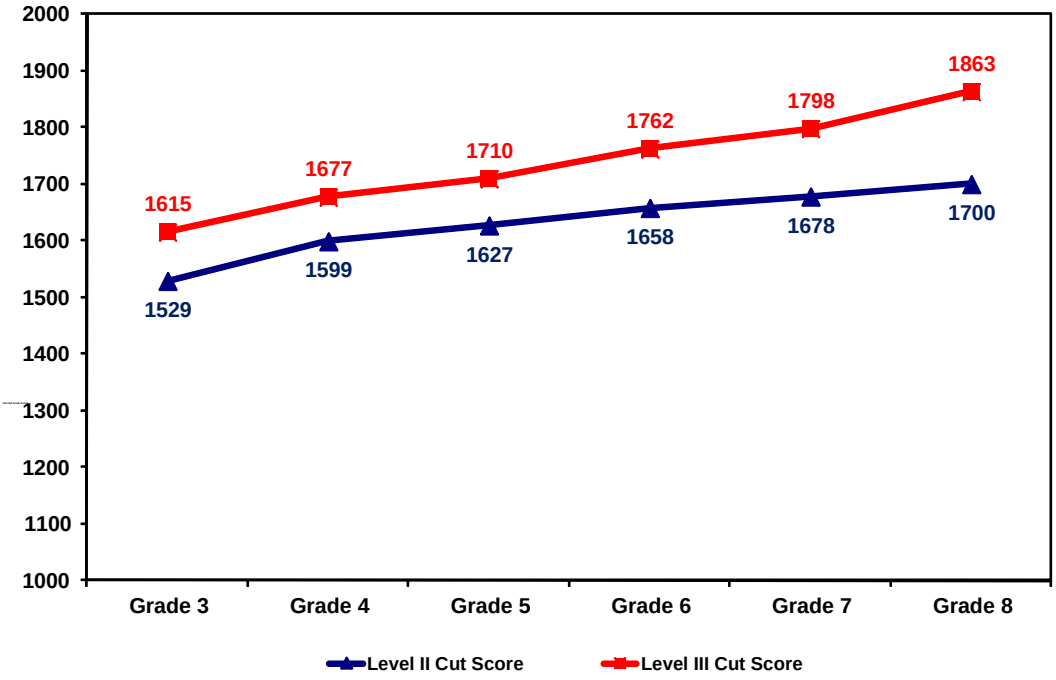


Figure 16. STAAR 3–8 Mathematics Final Recommended Cut Scores

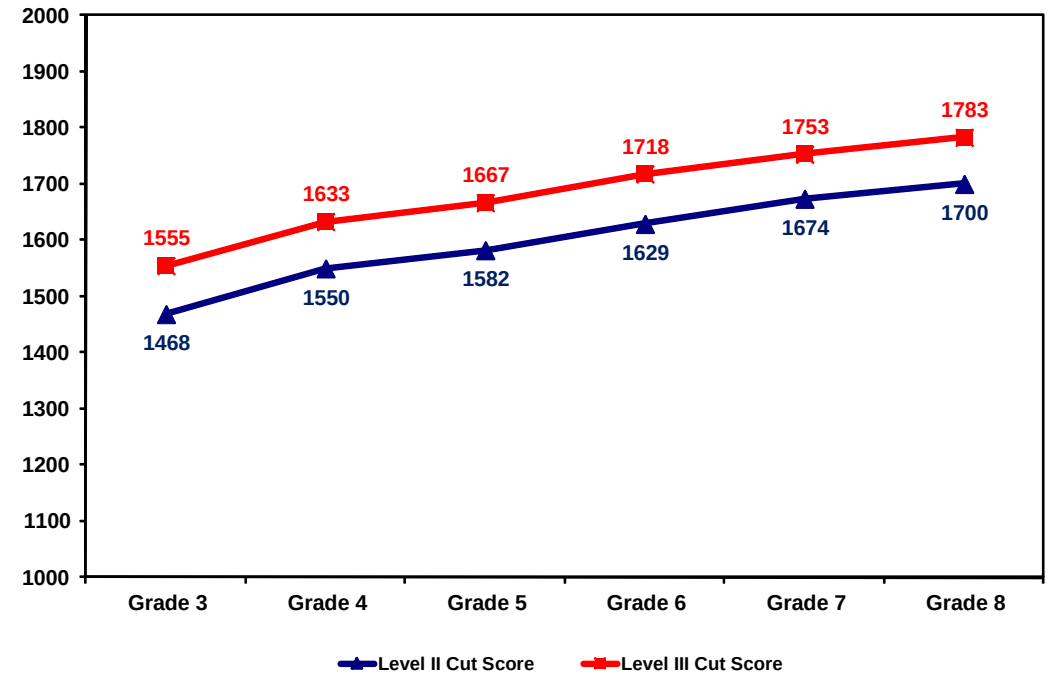


Figure 17. STAAR English 3–8 Reading Final Recommended Cut Scores

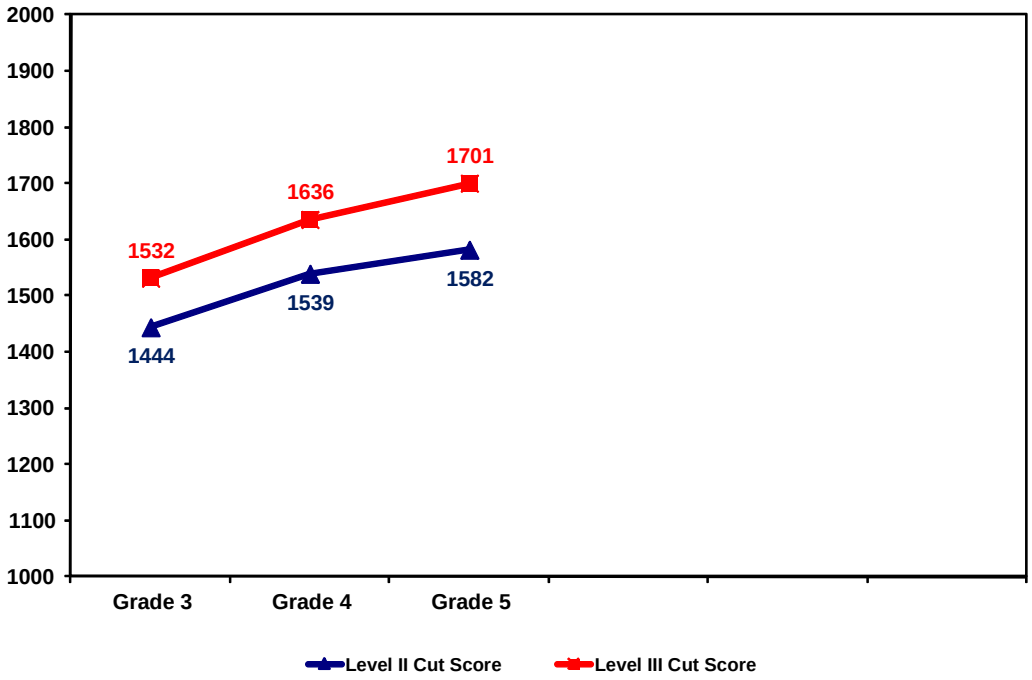


Figure 18. STAAR Spanish 3–5 Reading Final Recommended Cut Scores

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices* (2nd ed.). New York: Springer.

Linacre, J. M. (2001). *WINSTEPS Rasch Measurement Program, Version 3.32*. Chicago: John M. Linacre.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Patz, R. (2007). *Vertical scaling in standards-based educational assessment and accountability systems.* Paper presented at the Council of Chief state school officers.

Rasch, G. (1966). An Individualistic Approach to Item Analysis. In *Readings in Mathematical Social Science*, edited by Paul F. Lazarfeld and Neil W. Henry. Chicago, IL: Science Research Associates.

SAS/STAT (R) 9.22 User's Guide (n.d.) *Introduction to Statistical Modeling with SAS/STAT Software*. Retrieved May 23, 2013. http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_intromod_a0000000355.htm

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: Mesa Press.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: Mesa Press.

Appendix 1 – TTAC Notes

TTAC Discussion on STAAR 3–8 Vertical Scaling Bullets

Date: Monday April 25, 2011

Attendees: Michael Kolen (TTAC Member), Aimee Boyd, Sonya Powers, Malena McBride

- The English vertical scaling design is strong – having both lower grade level items and upper grade level items is ideal, especially because students will not receive scores on off-grade level items.
- If a choice must be made between upper and lower grade level items, include upper grade level items so the growth in content knowledge from before to after instruction can be measured.
- A minimum point-biserial of 0.2 is technically adequate and will allow more items with good content to be used on tests.
- Concurrent calibration within grades is a reasonable approach. Mean/Mean linking should be used to link across grades.
- To evaluate the reasonableness of using the Rasch model with the vertical scale, it is recommended that Rasch item fit statistics be evaluated. If many do not fit, check whether item discrimination changes across grades and if the 2PL model might provide more reasonable results.
- It is recommended that the English vertical scale design be used for Spanish too, if possible.
- 250 students per form for Spanish is okay as long as there are many vertical linking items used to calculate the vertical scaling constants.
- A hybrid design where vertical scale items are embedded in both field test and base test positions can be used to create a similar vertical scaling design for English and Spanish.
- With a hybrid design, item position effects are of concern. Keeping items in similar positions across grades is desirable.