2008 Spanish TAKS Vertical Scaling Study Report

November 13, 2008

Prepared for the Texas Education Agency
by
Pearson
Psychometric Services

Background

Tracking students' academic progress from year to year and over the course of their schooling has become increasingly important for large-scale assessment programs. The implementation of the No Child Left Behind Act of 2001 (NCLB; Public Law 107-110) requires annual assessment of student progress in the state curriculum standards at designated grade-based intervals. Student scores on such assessments are represented as numerical values that can be used to compute and describe differences in student performance within a given grade level. This scaling technique is termed *horizontal scaling*. Because horizontal scales are designed to represent within-grade level differences, they cannot be used to express student growth across grades. If the assessment of student performance requires the monitoring of student growth, then the expression of student performance requires the association of student scores with numerical values that reflect across-grade level performances. This scaling technique is called *vertical scaling*.

Since its inception in 2002-2003, the Texas Assessment of Knowledge and Skills (TAKS) has used a horizontal scale. However, under Section 39.036 in S. B. No. 1031, the Texas Educational Agency (TEA) is required to develop a vertical scale for assessing student performance beginning in the 2008-2009 school year in the following areas:

- mathematics, grades 3 through 8
- reading, grades 3 through 8

While no specific requirements were legislated for a vertical scale for the Spanish TAKS assessments, best practice would suggest that TEA evaluate the implementation of a vertical scale for the Spanish TAKS tests at grades 3-6. Therefore, data for a Spanish TAKS vertical scale score system were collected during spring 2008 so that a vertical scale score system could be in place by the 2009 administrations.

Data Collection Design

The data collection design for the Spanish TAKS vertical scaling study used embedded field-test positions on the assessments during the operational 2008 TAKS administrations. This design required designating either four or eight of the regular field-test forms for the vertical scaling study. Using the embedded field-test positions was desirable because students would have no knowledge of whether an item was an operational or vertical scale linking item.

The data collection model was based on a common-item nonequivalent group design with dual-grade common items for both the Spanish TAKS reading and mathematics tests. An example of this data collection is depicted in Figure 1 below. The model required using (a) only previously field-tested items as linking items, (b) both lower grade level and upper grade level linking items; and (c) embedded field-test positions. For example, some of the previously field-tested grade 3 items from the Spanish TAKS item bank were used as linking items and were placed on both the grade 3 and grade 4 vertical linking forms. Likewise, some of the previously field-tested grade 4 items from the item bank were used as linking items on grade 4, grade 3, and grade 5 vertical scale linking forms. The number of vertical scale linking forms needed for this model is eight for grades 4 and 5. Because the anchor items are only from one adjacent off-grade level, grades 3 and 6 only have four forms each.

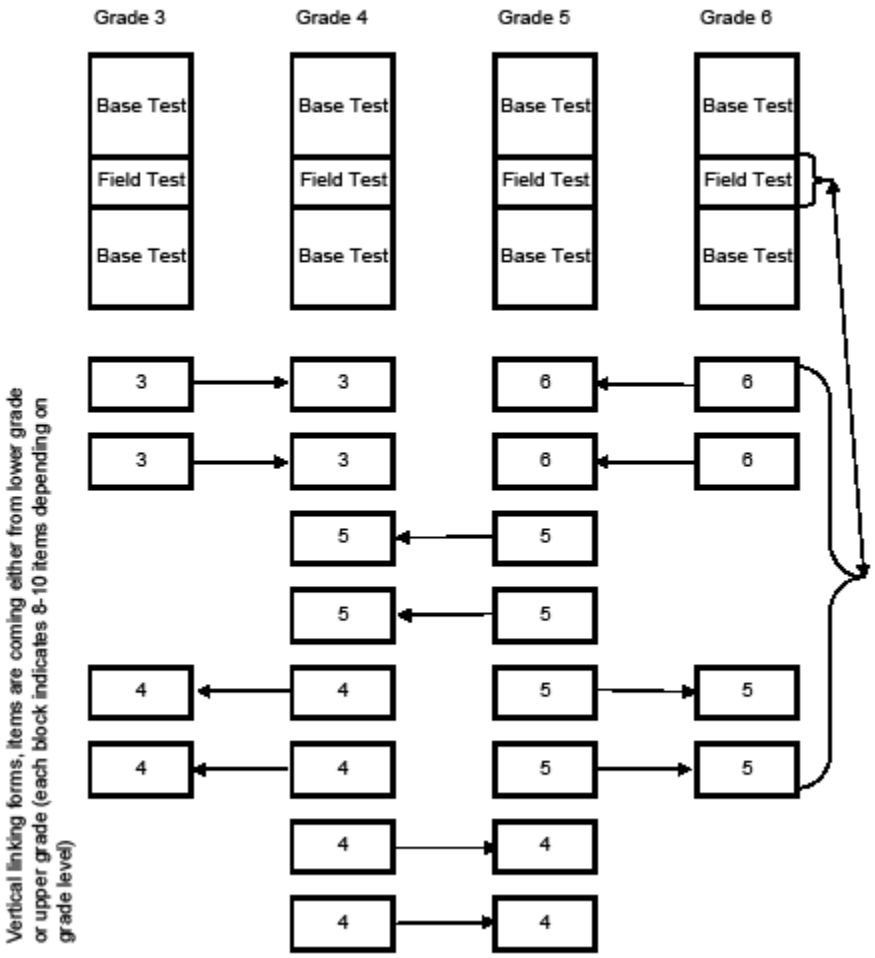


Figure 1. Spanish TAKS Reading and Mathematics Vertical Scale Data Collection Design.

Spanish TAKS Vertical Scaling Method

Spanish TAKS Vertical Scale Anchor Forms Descriptive Statistics

Vertical linking items used the same locations as field-test items. Table 1 displays the number of linking items between adjacent grade levels as well as the number of base test items for each grade level. Table 2 and Table 3 show the sample size for Spanish TAKS reading and mathematics vertical scale forms respectively. Note that the sample sizes for the Spanish vertical scaling study were especially limited at grades 5 and 6 where the total number of students in the Spanish testing population is small.

Table 1. Possible Number of Spanish TAKS Vertical Linking Items by Grade Levels

	Spanish TAKS Vertical Scale Grades				
	Direction	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**
Reading	Base Items	36	40	42	42
	Lower	--	20	20	20
	Upper	20	20	20	--
Math	Base Items	40	42	44	46
	Lower	--	16	16	16
	Upper	16	16	16	--

Table 2. Sample Size for Spanish TAKS Reading Vertical Scale Linking Forms

Form	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**
Overall	3936	7141	3599	944
1	--	--	--	307
2	993	873	436	211
3	984	885	450	211
4	993	908	486	215
5	966	906	429	--
6	--	896	433	--
7	--	888	419	--
8	--	893	459	--
9	--	892	487	--

Note: Spanish TAKS reading at grade 6 has 4 forms, and all forms were used for vertical scale.

Table 3. Sample Size for Spanish TAKS Mathematics Vertical Scale Linking Forms

Form	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**
Overall	3224	5797	2370	841
1	--	--	--	299
2	814	721	299	176
3	793	726	292	181
4	816	731	294	185
5	801	746	297	--
6	--	717	306	--
7	--	713	300	--
8	--	718	294	--
9	--	725	288	--

Note: Spanish TAKS mathematics at grade 6 has 4 forms, and all forms were used for vertical scale.

Main Steps for Developing Spanish TAKS Vertical Scales

1. Calibration of Vertical Linking Items

To develop a common scale across grade levels, vertical linking items were calibrated on their on-grade level and off-grade level with the live (i.e., operational) items. Based on feedback from the Texas Technical Advisory Committee (TTAC), within-grade calibration of anchor items and live items was done with separate (rather than concurrent) calibrations. The vertical linking study was conducted without linking to the TAKS base scale so that an estimate of across grade level changes in difficulty would not be contaminated by horizontal equating effects.

A two-step calibration procedure was used.

1) The first step was to calibrate only on-grade level live items together using all available students and to compute a mean ($\bar{b}_{On-grade\_only}$) of the resulting Rasch item difficulty estimates.
2) In the second step, the researcher calibrated both on-grade level live items and the vertical linking items (which includes both on-grade level and off-grade level items) together and computed an equating constant by computing the mean difference in on-grade level item parameters between the first step calibration ($\bar{b}_{On-grade\_only}$) and the second step calibration ($\bar{b}_{On-grade\_withVS}$) as shown in equation (1) below.

$$C = \bar{b}_{On-grade\_only} - \bar{b}_{On-grade\_withVS} \qquad (1)$$

This mean difference (C), the mean/mean equating constant (Kolen & Brennan, 2004), was applied to all item difficulties. This calibration method was followed form by form for all vertical linking forms.

2. Diagnostic Information

Model Misfit

Because the vertical scale is built using the Rasch model, it is important to detect any vertical linking item that does not fit this model. The Rasch Mean Square infit statistic was used to identity any misfiting items. If the Rasch Mean Square infit was unexpectedly large (>1.20) or small (< .08), then the item was eliminated from the vertical linking set at grade level and both adjacent grade levels (see Table 4 for more information).

Table 4. Items Eliminated During Model Misfit Analysis

Subject	Grade	Removed for Misfit
Spanish Reading	3-4	0
	4-5	0
	5-6	12
Spanish Math	3-4	3
	4-5	4
	5-6	9

Outliers

Another indicator of model misfit was outliers. A two-step process was used to identify outliers. First, a regression analysis was run using the on-grade level Rasch item difficulties (RID) and off-grade level RID. Although a visual inspection of regression plots, such as that shown in Figure 2, is a common method of outlier detection, it was decided to first use the externally studentized residual. This is the residual when the fitted regression excluding the current case under observation is divided by its error variance (which is also calculated excluding the current case under observation). An item was eliminated if the absolute value of its externally studentized residual was greater than 2.0, which was chosen because the studentized residual is distributed as a T- statistic. Second, the remaining Rasch item difficulties for the vertical linking items from on-grade and off-grade clusters were plotted against each other and then compared to a graph of items before any outlier was removed. The visual inspection process had two goals. The first was to verify that use of the externally studentized residuals was beneficial. Second, the visual inspection was used to eliminate any items that were missed by the externally studentized residual analysis but clearly needed to be removed. There were no additional items removed based on visual inspection.

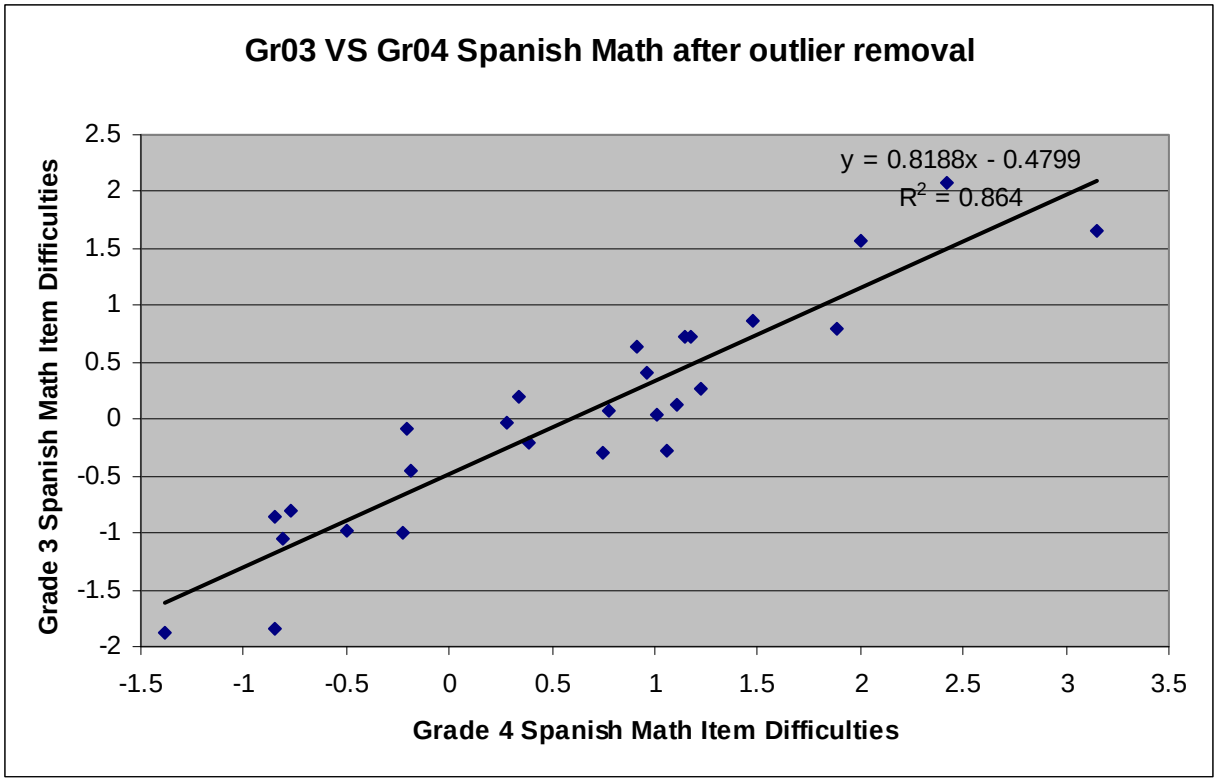


Figure2. Example Plot Used during Visual Inspection

Content Representation

It is not only important to inspect any vertical linking item set for items that are not functioning as expected, it is also important to review the removal of any vertical linking items in the context of their impact on the creation of the vertical scale. Table 5 provides an initial count of the vertical linking items by grade set, the number of items eliminated at each step of the diagnostic analyses, and the final number of items used to create the vertical scales.

Table 5. Spanish TAKS Eliminated Items due to Misfit and Outliers

Subject	Grade	# Vertical linking Items			Removed for Infit		Removed for outlier		# Vertical Linking items remaining		
		Total	Lower	Upper	Lower	Upper	Lower	Upper	Total	Lower	Upper
Reading	3-4	40	20	20	0	0	1	2	37	19	18
	4-5	40	20	20	0	0	1	2	37	19	18
	5-6	40	20	20	9	3	1	1	26	10	16
Math	3-4	32	16	16	1	2	1	1	27	14	13
	4-5	32	16	16	1	3	1	1	26	14	12
	5-6	32	16	16	3	6	2	0	21	11	10

Though 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 linking item set covers the content as was originally intended. Table 6 displays, by objective, the percentage of Spanish reading items for each adjacent grade set that were in the initial vertical linking item set, and the percentage in the final vertical linking item set after the removal of misfiting items and outliers.

Table 6. Spanish TAKS Reading Content Representation before and after Eliminating Items due to Misfit and Outliers

Subject	Grade	# Vertical Linking Items		Objective							
				1		2		3		4	
		Original	After	Original	After	Original	After	Original	After	Original	After
Spanish Reading	3-4	40	37	40%	41%	15%	17%	25%	22%	20%	22%
	4-5	40	37	33%	33%	10%	11%	30%	27%	28%	30%
	5-6	40	26	33%	31%	18%	19%	25%	27%	25%	23%

After reviewing the Spanish mathematics items, the final content representation of the mathematics grade 5-6 vertical linking item set did not match the initial content representation as well as we would like. Initially, all (n=4) the grade 5-6 common linking items were eliminated for objective 3, 'Geometry and Spatial Reasoning." One item was eliminated due to a Rasch Mean Square infit less than 0.80, while the other items were eliminated because their externally studentized residual was greater than 2.0. The addition of one of the eliminated items back into the vertical linking set would sufficiently satisfy the content representation for this objective. Furthermore, it was decided that given the use of the Rasch model, the addition of an item with an externally studentized residual greater than 2.0 was better than the addition of an item that did not fit the Rasch model. As a final verification, the item was selected that had least impact on the correlation between the grade 5 and grade 6 Rasch item difficulties for the Spanish mathematics grade 5-6 vertical linking item set before and after the inclusion of an item and was added back into the grade 5-6 vertical linking set for Spanish math. The correlation changed by .01. Table 7 displays, by objective, the percentage of Spanish mathematics items for each adjacent grade set that were in the initial vertical linking item set, and the percentage in the final vertical linking item set after the removal of misfiting items and outliers.

Table 7. Spanish TAKS Mathematics Content Representation before and after Eliminating Items due to Misfit and Outliers

Subject	Grade	# Vertical Linking Items		Objective											
				1		2		3		4		5		6	
		Original	After	Original	After	Original	After	Original	After	Original	After	Original	After	Original	After
Spanish Mathematics	3-4	32	27	16%	19%	13%	11%	19%	11%	16%	19%	16%	15%	22%	26%
	4-5	32	26	22%	19%	13%	12%	13%	12%	16%	19%	13%	12%	25%	27%
	5-6	32	20	22%	19%	19%	24%	13%	5%	16%	19%	13%	19%	19%	15%

3. Finding Vertical Linking Constants for Adjacent Grade Levels

The Mean/Mean method (Kolen & Brennan, 2004) was used to define vertical linking constants between the grade levels. The vertical equating constants were computed in three ways:

1) using only lower grade level vertical linking items
2) using only upper grade level vertical linking items
3) using a combined anchors set (both lower and upper grade level vertical linking items)

For example, the vertical linking constant between grades 4 and 5 would be the difference between the mean of the vertical linking Rasch item difficulties for grade 4 and grade 5. This process is depicted in Table 8. Table 9 provides the vertical linking constants for Spanish TAKS reading and mathematics.

Table 8. Calculation of Vertical Linking Constants for Spanish TAKS Reading and Mathematics

Grade	Vertical Linking Constant
3-4	$VC_{34}=VL_3-VL_4$
4-5	$VC_{45}=VL_4-VL_5$
5-6	$VC_{56}=VL_5-VL_6$

Table 9. Spanish TAKS Vertical Linking Constants by Linking Item Types.

Subject	Adjacent Grade Levels	Vertical Linking Constant Using All Items	Vertical Linking Constant Using Only Lower Items	Vertical Linking Constant Using Only Upper Items
Spanish Reading	3-4	0.82560	0.82133	0.83011
	4-5	0.76368	0.74587	0.78248
	5-6	0.51007	0.52939	0.49799

Subject	Adjacent Grade Levels	Vertical Linking Constant Using All Items	Vertical Linking Constant Using Only Lower Items	Vertical Linking Constant Using Only Upper Items
Spanish Mathematics	3-4	0.58904	0.44293	0.74639
	4-5	1.11483	1.04500	1.19631
	5-6	0.25852	0.21740	0.30375

Estimating vertical linking constants in three ways provided options to create the vertical scale in different ways. To quantify the difference in results between using the lower and upper grades vertical linking item sets for each subject, a Cohen's *d* was calculated. Because the number of vertical linking items sometimes differed between the lower and upper grades, a weighted, pooled average standard deviation was used. Cohen (1988) provided general guidelines for interpreting *d*. A *d* above an absolute value of .80 is considered large. A *d* between an absolute value of .50 and .79 is considered medium. A *d* between an absolute value of .20 and .49 is considered small. Table 10 provides the Cohen's *d* between the upper grade vertical linking constants and lower grade vertical linking constants for Spanish TAKS reading and mathematics. Only the differences in the vertical linking item sets for grade 3-4 and grade 4-5 mathematics would be considered to have a practical difference.

Table 10. Cohen's *d* between the Upper Grade Vertical Linking Constants and Lower Grade Vertical Linking Constants for Spanish TAKS Reading and Mathematics

Grade	Reading	Math
3-4	0.011	0.303
4-5	0.061	0.233
5-6	-0.050	0.140

4. Defining the Base Grade Level

The base grade refers to the anchoring point of the vertical scale and is a largely arbitrary decision. Because the base grade for the English TAKS vertical scale was chosen to be grade 8, the base grade for the Spanish TAKS vertical scale was chosen to also be the highest grade level provided for Spanish TAKS test administration, which is grade 6.

5. Computation of the Final Vertical Linking Constant

As indicated earlier, the Mean/Mean equating procedure was used to find the linking constants between adjacent grade levels. After finding the vertical linking constant between adjacent grades, a cumulative linking constant was defined from the base grade to the lower grade levels for any grade level that was not adjacent to the base grade. For example, at grade 4, the *vertical linking constant* (between grades 4 and 5) would be the difference between the mean of the vertical linking Rasch item difficulties for grade 4 and grade 5. On the other hand, the *cumulative vertical linking constant* (between grade 4 and the base grade level, grade 6), is the vertical linking constant between grades 5 and 6 *minus* the vertical linking constant between grades 4 and 5 (see Table 11 below). Tables 12 and 13 provided the cumulative vertical linking constants for Spanish TAKS reading and mathematics.

Table 11. Final Vertical Linking Constants for Spanish TAKS Reading and Mathematics

Grade	Final Vertical Linking Constant
3	$0 - VC_{34} - VC_{45} - VC_{56}$
4	$0 - VC_{56} - VC_{45}$
5	$0 - VC_{56}$
6	0

Table 12. Spanish TAKS Reading Final Vertical Linking Constants

Subject	Grade	Final Linking Constant Using All Items	Final Linking Constant Using Only Lower Items	Final Linking Constant Using Only Upper Items
Spanish Reading	3	-2.09935	-2.0966	-2.11058
Spanish Reading	4	-1.27375	-1.27526	-1.28047
Spanish Reading	5	-0.51007	-0.52939	-0.49799
Spanish Reading	6	0	0	0

Table 13. Spanish TAKS Math Final Vertical Linking Constants

Subject	Grade	Final Linking Constant Using All Items	Final Linking Constant Using Only Lower Items	Final Linking Constant Using Only Upper Items
Spanish Math	3	-1.9624	-1.70533	-2.24645
Spanish Math	4	-1.37335	-1.2624	-1.50006
Spanish Math	5	-0.25852	-0.21740	-0.30375
Spanish Math	6	0	0	0

6. Selection of the Final Vertical Scale

Spanish TAKS final vertical linking constants were plotted in Figure 3 (reading) and Figure 4 (math). Although all three methods produced similar results, the final vertical linking constants computed from all items will be more stable. As such the "all items" scale was recommended. The final vertical linking constants that result from the selected vertical scale would be used for future Spanish TAKS grade 3-6 mathematics and reading administrations.

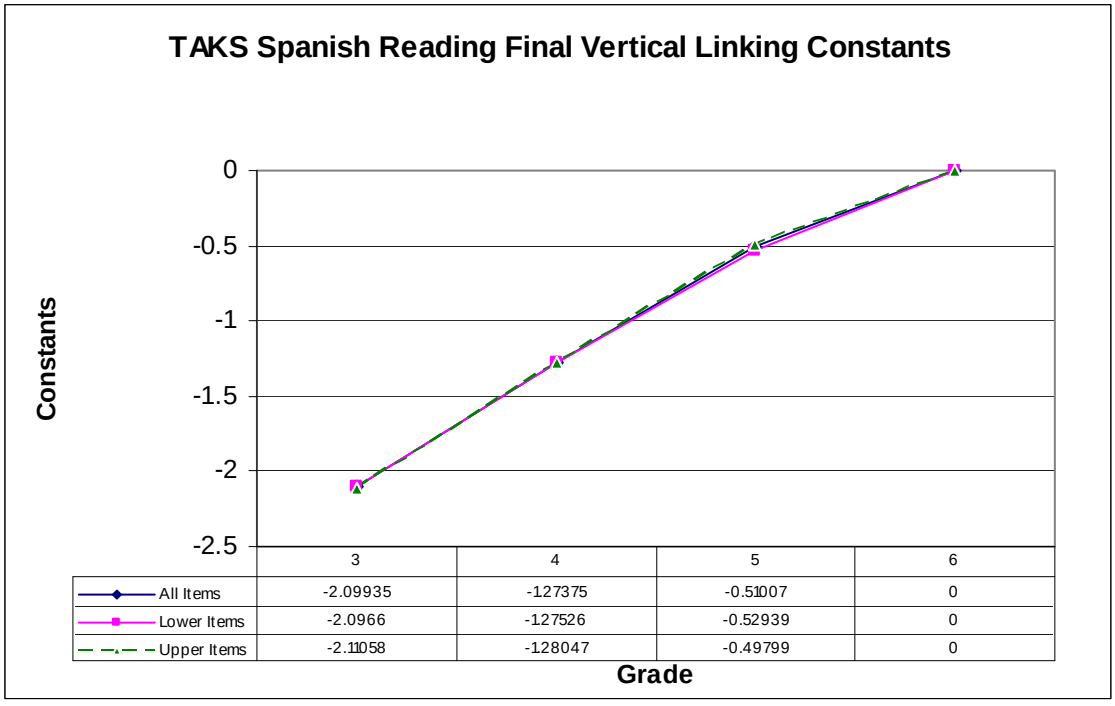
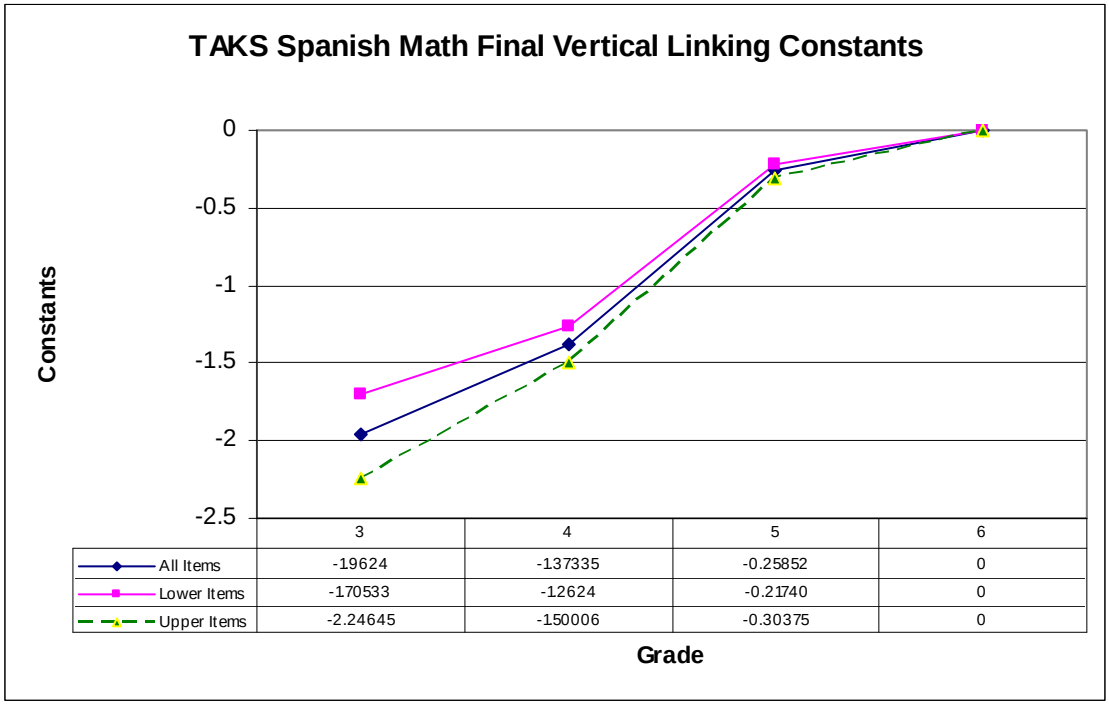


Figure 3. TAKS Spanish Reading Final Vertical Linking Constants



	3	4	5	6
All Items	-19624	-137335	-0.25852	0
Lower Items	-170533	-12624	-0.21740	0
Upper Items	-2.24645	-150006	-0.30375	0

Figure 4. TAKS Spanish Mathematics Final Vertical Linking Constants

7. Evaluation of Vertical Scale

The evaluation of a vertical scale is difficult. There are no firm guidelines by which one can judge all such scales. However, for the Spanish TAKS vertical scales, several criteria have been laid out for evaluating the scale. The development of these criteria was in part based on consultation with Dr. Michael Kolen and from a recent paper by Patz (2007). The following describe the Spanish TAKS vertical scale evaluation criteria:

1) Progression in difficulty across grades

- *Test-level progression*

It seems a reasonable assumption that, when comparing tests designed to assess 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 TAKS vertical scales is to verify that this progression exists. Figures 3 and 4 indicate that the vertical scales for both reading and mathematics progress show an upward trend, 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 grade 3 test.

- *Item-level Progression*

Another reasonableness check on a vertical scale is the performance of the vertical linking items. At the item level, it seems reasonable that an item should not perform dramatically differently, relative to the other items, across grades. One way to look at across grade differences is to examine the correlation coefficient between RIDs for items in a grade set. 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 14 provides the correlation between RIDs for the final adjacent grade common item sets. Across Spanish reading, the correlations were high, positive, and stable. Across Spanish mathematics, the correlation decreased from grade 3-4 (0.93) to grade 5-6 (0.77).

Table 14. Correlation of item RIDs for the adjacent level final vertical linking item sets for Spanish Reading and Mathematics

Grade	Spanish Reading	Spanish Math
3-4	0.98	0.93
4-5	0.92	0.87
5-6	0.93	0.77

2) Scale score means should increase across grade levels in a regular pattern.
Just as the vertical linking constants should progress in difficulty, it is reasonable to assume that the application of these linking constants should affect the population of interest in the same manner. Table 15 displays the means and standard deviations for the vertically scaled theta values for the scores of the students who participated in the vertical scaling study. Theta values were used because scale scores have not been assigned to the assessment yet. As expected, the means of the vertically scaled thetas increased across grade levels and the increase formed a regular pattern.

Table 15. Mean and Standard Deviation of Vertically Scaled Theta Values across Grade Levels for Spanish Reading and Mathematics

	Spanish Reading		Spanish Math	
Grade	Mean	SD	Mean	SD
3	-0.392	1.131	-0.068	1.300
4	0.160	1.161	0.274	1.249
5	0.759	1.062	0.611	1.062
6	1.301	1.038	1.271	0.942

3) Scale score SDs should not have large differences from grade to grade
As shown in Table 15, the standard deviations of the vertically scaled thetas are relatively stable with some decrease as the grade level increases.

4) The relationship between lower, upper, and combined vertical linking item sets should be regular (i.e., vertical scale line plots from each group do not cross) and the differences between the vertical scales derived from each vertical linking item set should be minimal (plot of final vertical linking constants).
Estimating vertical linking constants using three separate methods provides options to create the vertical scale in different ways. This also provides a good way to cross-validate the vertical scaling methodology. If the different vertical linking item sets provide the same estimate of growth, then the vertical linking constants should be very similar. If they provide different estimates of growth, then one would expect that differences would be approximately a constant. Figures 3 and 4 could be used to evaluate this criterion. For Spanish reading (Figure 3), it is easy to see that this criterion is met. The line graphs for the difference vertical scaling constants are virtually identical. For Spanish mathematics (Figure 4), there is a slight departure that begins in grade 5 and ends in grade 3.

6) The vertical scale should be reasonably close to the relationship between the grade level TEKS.

To evaluate this criterion, separate reading and mathematics meetings were held with content teams from TEA and Pearson. The meetings were broken into three segments: a brief overview of vertical scaling, discussion of the growth expectations given the TEKS and knowledge about the student population, and a review of the actual vertical scale for the subject area.

In discussion with Spanish reading content teams, the content expectations for the vertical scale matched well the results of the vertical scale. One exception was that for grades 5 and 6 Spanish reading, the content team expected less growth than what was actually observed.

For Spanish mathematics, the expectation was that the vertical scale should show fairly steady growth from grade 3-5. The growth would be less from grade 5 to 6 considering the instability of the grade 6 Spanish population caused by students who exit the bilingual program at grade 5. The expectation matched the results of the vertical scale results for Spanish mathematics.

The 2008 Spanish TAKS vertical scale will be implemented starting in spring 2009.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ. : Erlbaum.

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

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

Appendix A.

Vertical Linking Item Selection Guidelines

Vertical Linking Item Selection Guidelines
08/08/2007

Content Expectation of Vertical Linking Items

- Vertical linking items should have already been field tested.
- It is preferable to use the items that were field tested within the most recent three years.
- No griddable items should be used for the vertical linking.
- Choose items that do not require extra materials such as a ruler or calculator when used on an off grade level test.
- Vertical linking items should be content representative of the overlapping content from an adjacent grade.
- All of the vertical linking items from one grade level can be viewed as an on-grade level mini-test.
- Upper grade level vertical linking items should come from a content area where off grade students had exposure to the topic, especially in mathematics.
- Passage-based off grade level vertical linking items on a vertical linking form should come from one passage.
- Selected upper grade level passages should be as close as possible to the off-grade level requirements in terms of word counts etc.
- Vertical linking passages should reflect their grade level passage types.

Psychometric Properties of vertical linking items

- Avoid using extremely easy or difficult items, such that item difficulty (P-value) range should be .30 - .90.
- Rasch Item Difficulties should be within the -2.0 and +2.0 range; however, item difficulties around the mean of on grade level test are preferred.
- Choose items with high point biserials; point biserials should be greater than or equal to .30.
- Vertical linking items Rasch item fits should range from .80 to 1.20.