

# CHAPTER 13: SCALING

## Rationale

The basic score on any test is the raw score, which is the number of items correct, but the raw score alone does not present a broad picture of test performance because it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores should be used both to compensate for any differences in the difficulty of the items and to allow for direct comparisons of student performance between administrations.

## Rasch Partial-Credit Model

Test items (multiple-choice, gridded response, short-answer, and essay) for all Texas assessments are calibrated and equated using the Rasch Partial-Credit Model (RPCM). The RPCM is an extension of the Rasch one-parameter Item-Response Theory (IRT) model attributed to Georg Rasch (1980), as extended by Wright and Stone (1979), Masters (1982), Wright and Masters (1982), and Linacre and Wright (2001).

The RPCM was selected because of its flexibility in accommodating multiple-choice (correct/incorrect) data as well as multiple-response category data, and for its ability to maintain a one-to-one relationship between derived scores (that is, scale scores) and the raw scores. It is the underlying Rasch scale that facilitates equating of multiple test forms and allows for comparisons of student performance across years. Additionally, the underlying Rasch scale facilitates the critical maintenance of equivalent performance standards across years. The RPCM is defined via the following mathematical measurement model where, for a given item involving  $m$  score categories, the probability of person  $n$  scoring  $x$  on prompt  $i$  is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^x (B_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (B_n - D_{ij})},$$

where  $x = 0, 1, 2, \dots, m - 1$ , and

$$\sum_{j=0}^0 (B_n - D_{ij}) \equiv 0$$

The RPCM provides the probability of a student scoring  $x$  on the  $m$  steps of question/prompt  $i$  as a function of the student's proficiency level  $B_n$  (sometimes referred to as "ability") and the step difficulties ( $D_{ij}$ ) of the  $m$  steps in prompt  $i$ . (See Masters, 1982, for an example.) Note that for multiple-choice and gridded-response questions, 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 an item difficulty.

The application of the RPCM means that all multiple-choice items and open-ended tasks will be placed on the same scale. All common item- and step-difficulty estimates will be on the same underlying logit scale as that of the student proficiency level estimates. Estimates of items being field-tested can be obtained from a form-by-form or a concurrent calibration, with the common item set serving as an anchor. In this way, all field-test items can be placed on the same logistic scale as that of the common items.

At the conclusion of these calibrations, all item- and task-difficulty estimates as well as all student proficiency level estimates are directly comparable because they are on the same underlying logistic scale.

The advantages of an RPCM scaling include:

- All items, independent of type, are placed on the same common score scale.
- The RPCM provides the same score scale onto which students' achievement results are placed. Hence, direct comparisons regarding the kinds of items students with various achievement levels can answer can be made. This is very helpful in describing the results of the assessments to students, parents, and teachers.
- All field-test items can be placed on the same scale as those of the live, or operational, part of the assessment. This is invaluable in linking student performance back to all banked items and useful in the construction of multiple future forms that are psychometrically balanced.
- This design allows for the pre-equating of future test forms, which is a valuable component of the complex test construction process.
- Such an approach supports post-equating of the test. In this way, a link is established between previous forms and the current administration. This current form is on the same scale as the previous forms so that comparisons in form difficulties and passing rates can be ascertained. Because both pre-equated and post-equated item difficulty estimates are available, any difficulty drift or scale drift can also be quantified.
- Establishing a common scaling allows for the direct comparison of performance-level standards established by the SBOE against future test forms.

# TAKS

## Scale Scores

Although the RPCM model provides the underlying scale for the TAKS tests, it is not a metric that is useful for reporting purposes due to the properties of that scale. For the TAKS tests, scale scores have been developed through a linear transformation of the underlying Rasch proficiency (“ability”) estimates to ensure that the performance standards are maintained at the same level of difficulty across administrations. The TAKS scale scores ( $SS$ ), which are derived scores, are useful in describing different aspects of student performance and maintaining performance standards over test administrations.

Derived scores are computed using the one-parameter IRT or RPCM. The advantage of using IRT models in scaling is that all the items measuring performance in a particular content area can be placed on the same scale of difficulty. The further value of the Rasch model over more complex IRT models is that the Rasch model assumes that for each total score point, there is only one student proficiency (or ability) estimate. This relationship allows the Rasch difficulty values for individual items to be used in computing a Rasch ability level for any total score point on any test constructed from these items.

The SBOE established the performance standards for most TAKS tests independently at each grade level and test content area in November 2002. During the spring 2003 operational test administration, tests were initially calibrated onto a Rasch partial-credit model scale. For TAKS developed since 2002, the SBOE established performance standards and the initial calibrations onto a Rasch partial-credit model scale have been conducted. Calibration of the TAKS operational test data was accomplished by Pearson Educational Measurement with independent verification of the analyses performed by TEA and an external psychometric consultant. The extensive verification procedure was part of a TEA quality assurance plan that was put into place to ensure the accuracy of the results of the Rasch partial-credit scaling of TAKS.

A unique scale transformation was then developed in each grade and content area so that the resulting set of scale scores would have the panel-recommended Met Standard performance-level cut set at a scale score of 2100 and the panel-recommended Commended Performance-level cut set at a scale score of 2400. It was felt that establishing the panel-recommended cut scale scores to have the same value, regardless of test or grade level, would aid in the interpretability of the scale scores. This linear transformation of the underlying Rasch proficiency level estimate is as follows:

$$SS_j = (\theta_j \times T1) + T2$$

where  $SS_j$  is the scale score for student  $j$ ,  $\theta_j$  is the Rasch partial credit model proficiency level estimate for student  $j$ , and  $T1$  and  $T2$  are scale score transformation constants that establish the scale score system such that a scale score of 2100 is the cut score for the Met Standard performance level and a scale score of 2400 is the cut score for the Commended Performance level. Values for  $T1$  and  $T2$  are provided in Tables 13 and 14.

These linear transformations established the original scale score system based on the Rasch partial credit scaling of the spring 2003 test results. Statistical equating has been applied to maintain the same level of difficulty for newly developed forms.

The resulting TAKS scale score system has a range of approximately 1000 to 3200. For tests containing constructed-response items (open-ended or essay questions), it is important to note that the total score is a combination of the number-correct score on the multiple-choice questions and the number of points achieved on the constructed-response questions.

For the grade 10 and exit level English language arts tests, the total-score-to-Rasch-proficiency-level-estimate (and subsequent scale score) table incorporates a weighted essay score (essay score  $\times$  4). This helps ensure that the appropriate value is placed on the direct writing sample given the amount of time and effort put into writing to the essay prompt. Thus, for ELA it is important to note that the total number of attainable score points after weighting is greater than the number of items.

Additionally, scale scores for writing and ELA are impacted by the essay score requirement of the standards. For writing and ELA a student is required to attain a score of 2 or higher on the essay prompt in order to have Met Standard. For writing a student is required to attain an essay score of 3 or higher on the essay prompt in order to achieve Commended Performance. For additional information about the essay score requirements of the standards, see the Student Assessment Division website at [www.tea.state.tx.us/student.assessment/taks/standards/scalescorecuts0305.pdf](http://www.tea.state.tx.us/student.assessment/taks/standards/scalescorecuts0305.pdf).

If a student receives a score of '0' or '1' on the essay prompt, the highest scale score that he or she can receive is one scale score point less than Met Standard. For example, at the panel-recommended standard of 2100, the highest scale score a student can receive if he or she scores below a 2 on the essay prompt is 2099. All students receiving a '0' or '1' on the essay prompt with scale scores higher than this value as obtained through the Rasch calibration have their scores artificially "re-mapped" to this value to reflect the essay score requirement of the passing standard. Similarly, for writing the highest scale score a student can receive if he or she scores below a 3 on the essay prompt is 2399 (one scale score point less than the Commended Performance standard of 2400). Students with scale scores above this value based on the Rasch calibration will have their scores re-mapped. This can be observed as a spike (large number of students) at the re-mapped value in the scale score distribution in Appendix 24.

**Table 14. Scale Score Transformation Constants for the TAKS Tests (English)**

<b>English</b>	<b>T1</b>	<b>T2</b>
Gr 3 Reading	125.89173	1992.23668
Gr 3 Mathematics	146.69927	1967.23716
Gr 4 Reading	129.42192	1996.07420
Gr 4 Mathematics	142.51781	1976.29454
Gr 4 Writing	110.88114	1981.33501
Gr 5 Reading	155.92516	1954.52183
Gr 5 Mathematics	170.35775	1939.18228
Gr 5 Science	187.96992	1832.51880
Gr 6 Reading	166.38935	1988.85191
Gr 6 Mathematics	174.31726	1987.91400
Gr 7 Reading	139.08206	1964.53408
Gr 7 Mathematics	133.98839	2046.53863
Gr 7 Writing	135.59322	2002.82034
Gr 8 Reading	153.76730	1948.53921
Gr 8 Mathematics	153.68852	2025.61475
Gr 8 Social Studies	145.41929	2085.16723
Gr 9 Reading	123.21847	1944.27650
Gr 9 Mathematics	184.61538	2009.90769
Gr 10 ELA	97.06539	1983.74478
Gr 10 Mathematics	141.04372	2038.64598
Gr 10 Science	160.42781	1996.84492
Gr 10 Social Studies	145.20813	2046.85382
Gr 11 ELA	113.48162	2017.62369
Gr 11 Mathematics	140.58107	2064.71415
Gr 11 Science	129.47777	2070.86750
Gr 11 Social Studies	126.47555	2093.29680

**Table 15. Scale Score Transformation Constants for the TAKS Tests (Spanish version)**

<b>Spanish</b>	<b>T1</b>	<b>T2</b>
Grade 3 Reading	148.66204	1995.19326
Grade 3 Mathematics	146.69927	1968.26406
Grade 4 Reading	165.10732	2006.21904
Grade 4 Mathematics	198.15059	1923.64597
Grade 4 Writing	151.04980	1998.90237
Grade 5 Reading	190.23462	1967.02600
Grade 5 Mathematics	190.71837	1915.95677
Grade 5 Science	189.27455	1841.07256
Grade 6 Reading	187.96992	2057.89474
Grade 6 Mathematics	202.56583	1970.76300

Following the spring 2003 operational test calibration analyses, these linear transformations were applied to the resulting Rasch student proficiency (ability) estimates at each total score point, yielding the final raw score to scale score conversion tables, as shown in Appendix 25.

## Vertical Linking

TAKS is a standards-referenced assessment reflecting the curriculum as specified in the TEKS at each grade level. As part of the process for setting standards for student performance, groups of Texas educators participated in advising the SBOE on a recommended score point for each subject area at which students are assumed to have sufficient mastery of the TEKS student expectations at that grade level. The TAKS scale score system was set such that a scale score of 2100 is the minimum attainable panel-recommended Met Standard score and a scale score of 2400 is the minimum attainable Commended Performance score at each grade level and for each content area (though there are additional requirements for ELA). It was argued that such a scoring system, once the standards phase-in was completed, would be easier for students, parents, schools, and the public to understand since the meaning of a scale score (SS) of 2100 and 2400 would remain the same regardless of grade and subject.

These scales are grade and subject specific; however, they cannot be compared across years. Unlike TAAS, TAKS has no vertical scale score system, no Texas Learning Index (TLI), and no measure of student-level growth from grade to grade. The Texas Growth Index (TGI) provides a measure of growth; however, the growth is intended only for interpretation at aggregate educational units, such as campuses and districts. The TGI is calculated at the student level, but the reliability of student-level growth is not strong enough for interpretation at the student level. The TGI may add value to campuses struggling to show the movement of students within and across the proficiency levels. For more information on the TGI, see Chapter 9. The following discussion provides background concerning why there is no vertical scale underlying TAKS achievement assessments.

## Reasons Why No Vertical Scale Underlies TAKS

Because of differences in content measured across the grades, it is not technically defensible to formally equate, or link, the grades 3–8 scale with the grades 9–11 scale. For example, it would not be reasonable to equate general mathematics as measured in elementary or middle school with algebra and geometry as measured on high school TAKS. Vertical linking or equating requires the same construct to be measured from grade to grade. The national Technical Advisory Committee believed that one vertical linking system was needed across all grades, which could not be achieved due to the changing constructs being measured. For similar content reasons, it is not possible to establish a TLI-like metric on the exit level TAKS and then apply it at lower grades. The changing construct renders such a TLI-like statistic uninterpretable.

Performance standards were established separately for each grade and subject; therefore, these standards would likely not reflect the same performance expectations each year, if placed on a vertical scale score system. For example, under a vertical scale score system, the passing standard for grade 7 reading might represent a lower achievement expectation than that set for the grade 6 reading assessment. Though this may be appropriate from an educational perspective, as it might be reasonable to expect higher performance in some years than others, this would possibly cause confusion for users of the assessment system.

The vertical scale (as well as a TLI-like measure) is an abstract construct, one that cannot be interpreted directly. Some members of the national TAC voiced concerns that an incremental vertical pattern of achievement was not a reasonable representation of educational growth. They made a philosophical argument (more than a psychometric one) that students learn at different rates and at different times and that the vertical scale was an artificial representation of a much more complex growth pattern. One example of this can be seen in an artifact typical of vertical scales: a student answering no questions correctly at grade 3 in one year and no questions correctly in grade 4 of the following year could still show growth because the vertical scale score the second year associated with an earned zero score could be higher than the previous year.

One of the primary disadvantages of any derived score system (including vertically linked scale scores) is that it may lead to misinterpretation and confusion about patterns of score growth. For example, suppose the vertical scale spans grades 3–8 mathematics. If a student in grade 3 gets almost every item correct, he or she may attain a vertical scale score that is at, or above, the Met Standard scale score for the grade 5 mathematics test, even though the student has not yet had the benefit of fourth- and fifth-grade instruction. As another example, as a result of differences in test difficulty from year to year, a student who answers 10 questions correctly out of 40 this year might receive a higher scale score than he or she did the previous year, when he or she answered 12 questions correctly out of 40. This potential for misinterpretation was another of the reasons why some members of the national TAC did not support the concept of a vertical scale.

Under the rules of the No Child Left Behind regulations, an individual student growth measure is not required. NCLB primarily uses the percentage of students classified at each performance standard (in addition to other elements, such as participation rates) to measure Adequate Yearly Progress (AYP) used in most accountability programs. As such, the need for a student growth measure was not pressing for federal accountability reasons and was not recommended by the national TAC.

## **Algebra I End-of-Course**

### **Scale Scores**

TEA established the new performance standards for Algebra I End-of-Course in November 2005, replacing the old performance standards used in 2003 and 2004, which only had one cut scale score (at 1500) and two performance levels. Using a procedure similar to TAKS, a unique scale transformation was developed so that the resulting set of scale scores would have the panel-recommended Met Standard performance-level cut set at a scale score of 1100 and the panel-recommended Commended Performance-level cut set at a scale score of 1400. This linear transformation of the underlying Rasch proficiency level estimate is as follows:

$$SS_j = (j \times \mathbf{T1}) + \mathbf{T2}$$

where  $SS_j$  is the scale score for student  $j$ ,  $ij$  is the Rasch model proficiency level estimate for student  $j$ , and  $T1$  and  $T2$  are scale score transformation constants that establish the scale score system such that a scale score of 1000 is the cut score for the Met Standard performance level and a scale score of 1400 is the cut score for the Commended Performance level. Values for  $T1$  and  $T2$  are 155.0468 and 1009.0186, respectively.

This linear transformation established the original scale score system based on the Rasch dichotomous scaling of the spring 2005 test results. Since the new standards were set after the reporting of spring 2005 administration results, the performance levels for students were not provided for the students tested in that administration; instead, the test report contained their raw scores. Fall 2005 was the first time the performance standards were used with the new Algebra I End-of-Course test scores.

## **SDAA II**

Unlike its predecessor, SDAA, the SDAA II program does not have a vertical scale. Reasons for this decision are similar to those presented for TAKS in the TAKS: Vertical Linking section of this chapter. The SDAA II to TAKS Linking Study Report is included as Appendix 22. SDAA II now measures student achievement at both the elementary/middle (grades 3–8) and high school (grades 9–10) levels. After consultation with the Texas Technical Advisory Committee (TTAC), TEA and PEM concluded that due to differences in the content measured across these grades, a formal vertical equating or linking of the tests would not be technically defensible. Instead, student growth on SDAA II can be tracked by evaluating progress through instructional and achievement levels (see below) from one year to the next.

As with TAKS, SDAA II uses a Rasch partial-credit IRT model to determine the difficulty of items within each instructional level and content area and to equate test forms across time. As with TAKS, the underlying Rasch scale does not have desirable properties for reporting purposes. Instead of using a linear transformation of the underlying scale to report scale scores as is done with TAKS, SDAA II scores are reported using raw scores and Achievement Levels. This decision was made in consultation with the TTAC to aid in the usability and interpretation of scores reported from the SDAA II system. The advice of the TTAC was to focus interpretation of SDAA II scores on the Achievement Levels attained by students. Though the underlying Rasch scale will still be used to equate tests from year to year, the raw scores will be used for reporting purposes.

There are three achievement levels within each SDAA II instructional level test. Each achievement level spans a range of scores, with no items correct at the beginning of Achievement Level I and all items correct at the end of Achievement Level III.

The achievement level in SDAA II has two purposes: (1) it describes a student's performance on the SDAA II in a manner that is meaningful to students, parents, and members of the ARD committee and (2) it allows for an evaluation of a student's progress from year to year. The achievement levels are broad performance categories that minimize some of the technical difficulties presented by measures of growth, particularly issues associated with test reliability. A student's achievement level is determined by the number of items he or she answers

correctly. The raw score required to be categorized within an achievement level may change slightly from year to year due to equating. A description of the performance associated with each SDAA II achievement level follows.

**Level I:** A student scoring at this achievement level demonstrates **minimal** knowledge and skills related to the TEKS student expectations at the appropriate instructional level for reading and/or mathematics. Performance at this level indicates that the student is considered to be **beginning** and has an understanding of **few** of the required concepts in the reading and/or mathematics TEKS.

**Level II:** A student scoring at this achievement level demonstrates **adequate** knowledge and skills related to the TEKS student expectations at the appropriate instructional level for reading and/or mathematics. Performance at this level indicates that the student is considered to be **developing** and has an understanding of **some** of the required concepts in the reading and/or mathematics TEKS.

**Level III:** A student scoring at this achievement level demonstrates **strong** knowledge and skills related to the TEKS student expectations at the appropriate instructional level for reading and/or mathematics. Performance at this level indicates that the student is considered to be **proficient** and has an understanding of **most or all** of the required concepts in the reading and/or mathematics TEKS.

The reported SDAA II achievement level notation combines the assessed instructional level with the demonstrated achievement level. For example, an achievement level of 2–I means the student was assessed at Instructional Level 2 and demonstrated minimal knowledge and skills (Level I).

Achievement level cuts for reading and mathematics were established in spring 2005 on the raw score by a panel of content experts. These raw scores were then translated to the Rasch scale to permit equating of the achievement levels within each test from year to year. The resulting Rasch scale cuts and range of raw scores associated with each of the achievement levels is shown in Table 2 in Chapter 3. For further information on setting the achievement level cuts, see Appendix 18 in the 2004–2005 *Technical Digest*. The underlying Rasch scale will control for variations in overall test difficulty from year to year with the result that the raw score necessary to be classified within a given achievement level may change slightly each year.

Achievement levels for the SDAA II writing tests for Instructional Levels 2 through 8/9 and SDAA II English language arts Instructional Level 10 were assigned based on a combination of scores from the non-essay portion of the writing test (that is, multiple-choice and open-ended item portion) in conjunction with scores from the essay portion of the writing test. For more information on how to determine achievement levels for the SDAA II writing test, see Chapter 17: Performance Assessment. Additional information about the SDAA II can be found in Chapter 3: SDAA II.

## RPTE

Rather than linking performance to grade-level expectations in the traditional sense, RPTE measures performance in terms of language proficiency levels that describe what second language learners can read and understand at various stages of English acquisition. RPTE reports performance at four language proficiency levels—beginning, intermediate, advanced, and advanced high. Students who enter U.S. schools knowing no English, regardless of their grade level at the time of entry, progress from one proficiency level to the next as they become fluent in English.

Both TAKS and RPTE measure the reading skills required by the TEKS. However, the manner in which RPTE assesses these skills reflects the stages of second language acquisition, which occur on a continuum spanning from little or no knowledge of English to full English fluency. RPTE tests are constructed as mini-tests within a test. Each English language learner takes the entire assessment. The English used on each mini-test is appropriate for students at that stage of English acquisition. The student's performance on RPTE is reported in terms of proficiency levels. The proficiency level rating a student receives on RPTE indicates the highest English proficiency level at which the student reads with understanding and performs TEKS reading skills successfully. RPTE, in essence, is a linguistically accommodated assessment. Its built-in accommodations, linked to stages of second language acquisition, allow the assessment to provide useful information about both the development of a student's reading skills and his or her progress in learning English. The English reading proficiency of LEP students is expected to increase annually as they continue learning English and receiving instruction in reading.

Given this background, it is not surprising that the RPTE test is based on an underlying Rasch vertical scale that allows the test results to track individual student progress in English proficiency over time. Vertical scaling of RPTE was obtained via a fall 2000 RPTE scaling study. Special hybrid test booklets were constructed so that each test booklet contained a mini-form of items from each of two adjacent RPTE grade-level clusters. Students took a mini-form that contained items appropriate for their enrolled grade level as well as items that were one grade cluster lower. This type of design is very common in vertical linking studies in educational testing.

Using grade 3 as the base form (for vertical linking purposes), the vertical linking constants are cumulative across the RPTE grade clusters with respect to scale distance from the base form scale (in this case grade 3). These cumulative vertical linking constants (referred to as the scaling constants) are equal to the vertically scaled mean item difficulty of the test items at each grade cluster. Since grade 3 was used as the base (and a scaling constant of zero), it would have a scaled mean item difficulty value of zero.

**Table 16. Vertical Scaling Constants for RPTE**

<b>Grade Cluster</b>	<b>Constant</b>
3	0.000
4–5	0.680
6–8	0.772
9–12	1.562

As with TAKS, the underlying Rasch scale does not have desirable properties for reporting purposes, so the final student proficiency (ability) levels are subjected to a linear transformation in order to derive the RPTE scale scores. As can be seen from the vertical scaling constants, the RPTE vertical scale is centered on the grade 3 cluster as the zero point. The RPTE scale score transformation is as follows:

$$SS_j = (\theta_j \times 48) + 575$$

where  $\theta_j$  is the vertically scaled Rasch student proficiency level for student  $j$ .

This scale score system results in scale scores in the range of 300 to 900 and is flexible enough for development of future test forms. RPTE student proficiency levels based on this scale score metric were then determined based on the spring 2000 test, using the final raw score proficiency-level cuts.

**Table 17. Scale Score Ranges Associated with RPTE Proficiency Levels**

<b>RPTE Grade Cluster</b>	<b>Beginning SS Range</b>	<b>Intermediate SS Range</b>	<b>Advanced SS Range</b>	<b>Advanced High</b>
3	623 and below	624–685	686–734	735 and above
4–5	657 and below	658–719	720–789	790 and above
6–8	666 and below	667–720	721–814	815 and above
9–12	692 and below	693–743	744–829	830 and above

## Proficiency Level Descriptors

Descriptors of the four RPTE reading proficiency levels are found on pages 108–110. Additional information about RPTE can be found in Chapter 4: TELPAS.

## Quality Control

The scaling process for TAKS, SDAA II, and RPTE is independently conducted by at least two different psychometricians at Pearson Educational Measurement. Once each party completes

the Rasch calibrations and applies the scaling transformation (where appropriate), the separate results are compiled. These compiled results are reviewed for differences. If any differences are detected, the results and procedures are reviewed until consensus is reached.

## TAAS Exit Level

The TAAS testing program is still ongoing, in retest form, for students for whom TAAS is their high school graduation requirement. TAAS provided two derived scores that described different aspects of student performance: scale scores and Texas Learning Index (TLI) scores.

### Scale Scores

For the TAAS tests, scale scores were developed to ensure that the performance standards were maintained at the same level of difficulty across administrations. The original test form, on which the 70 percent correct standard was established (see Chapter 12: Standards), was calibrated using the Rasch model. This calibration produced a relationship between the raw score and the Rasch achievement score, which was then transformed so that 1500 represented the passing standard. This transformation is reproduced in the formula

$$\text{TAAS Scale Score} = \left( \frac{\theta - \theta_{\text{at Standard}}}{\sigma_{\theta}} \right) \times 200 + 1500$$

where  $\theta$  is the Rasch student proficiency level estimate,  $\theta_{\text{at Standard}}$  is the Rasch student proficiency level associated with the 70 percent raw score,  $\sigma_{\theta}$  is the standard deviation of the Rasch student proficiency levels, and 200 and 1500 are the spread and centering constants, respectively. This transformation established the original scale score system, and statistical equating was required to maintain the same level of difficulty for newly developed forms. The resulting TAAS scale has a range of approximately 400 to 2400, with 1500 corresponding to 70 percent of the items correct on the first administration of the test, when the passing standards were set.

### Texas Learning Index

The Texas Learning Index (TLI) was developed to better meet the needs of districts and students for longitudinal comparability. A metric with two essential characteristics was sought. First, such a metric should provide an index of student achievement toward the goal of passing the TAAS exit level test. Second, the metric should permit comparisons between administrations and between grades for use in the accountability system. The TLI provides a means for schools to be able to demonstrate improvements in their instructional programs even in cases where the passing standard has not yet been met or the passing standard has been exceeded. Likewise, with a derived score such as the TLI, individual students are able to demonstrate improvement regardless of their current achievement relative to the passing standard.

## T-Score Type Transformation

The requirements listed above led to the consideration of a vertical scale score system for the TAAS examinations that would place results for grades 3–8 and the exit level test all on the same scale. TEA convened a panel of measurement experts from across the nation to advise the agency regarding such a scaling. This panel was composed of educators, test publishers, and educational consultants. The committee expressed two main concerns regarding a vertical scaling system. First, placing both grade 3 and exit level students on the same scale could lead to misinterpretations because of the large difference in the content of the test items at these grades. Second, a vertical scale implies a linear and well-defined curriculum from grades 3 through exit level when such a well-ordered curriculum may not be in place. The committee concluded that a vertical scale would not meet the needs of TEA and offered an alternative proposal of using a transformed “T-score” type of scale. A transformed T-score is expressed in terms of standard deviation units away from the mean. For example, if a student earns a raw score of 50 on a test with a mean of 40 and a standard deviation of 10, this student’s score is one standard deviation unit above the mean. Traditionally, such a score is referred to as a standard score, or z-score, and can be reproduced with the following formula.

$$z = \frac{\text{(Observed Score – Mean)}}{\text{Standard Deviation}}$$

The student in the example above has an observed score of 50 and a z-score of 1.00. Because standard scores have decimals and typically range from -3.00 to 3.00, an additional transformation is usually made to simplify the reporting. A common transformation in which the scale in this example is “re-anchored” to have a mean of 50 and a standard deviation of 10 is often referred to as a T-score transformation. The following formula provides the T-score transformation.

$$T = (z \times 10) + 50$$

Such a transformation simply renames the more-difficult-to-use z-score. After the transformation, the student in the example would have the following scores: a raw score of 50, a z-score of 1.00, and a T-score of 60.

## TLI Established at Exit Level

The TLI is very much like the T-score previously described. Unlike the T-score, however, the TLI is anchored at the exit level passing standard rather than at the mean of the distribution. To distinguish between the scale score system and the TLI, TEA chose a two-digit metric for the TLI so that it is anchored at the exit level passing standard with a value of 70 and a standard deviation of 15. The TLI is derived by the following formula.

$$TLI = \left[ \left( \left( \frac{\text{(Observed Score – Mean)}}{\text{Standard Deviation}} \right) - \text{z-Score at Passing Standard} \right) \cdot 15 \right] + 70$$

Because a TLI of 70 represents the passing standard, there is no difference in interpretation between a student who scores 1500 under the scale score system and a student who scores 70 on the TLI scale. The TLI was first established in spring 1994.

## **Assumptions**

An assumption of the TLI is that the reference distributions on which the TLI scale was constructed should be used for all future TAAS scoring. Because the TLI is a distributional-based metric relying on a z-score transformation, it is normative in nature. Because of this norm-referenced component, recalculating the TLI each year would make year-to-year comparisons impossible. For this reason, all TAAS administrations describe student performance in terms of the population tested in spring 1994.

Additional information about the TAAS scaling, including the TLI derived score, can be found in the 2001–2002 *Technical Digest*.

## **Frequency Distributions and Descriptive Statistics**

Appendix 24 provides frequency distributions and summary statistics for the TAKS and RPTE scale scores and the SDAA II raw scores. Appendix 26 provides mean p-values by objective and subject area and internal consistency estimates for TAKS, RPTE, and SDAA II tests taken.