EDUCATIONAL MEASUREMENT:
ISSUES AND PRACTICE

NCME

Educational Measurement: Issues and Practice
Winter 2010, Vol. 29, No. 4, pp. 39–48

Scaling: An ITEMS Module

Ye Tong, *Pearson*, and Michael J. Kolen, *The University of Iowa*

Scaling *is the process of constructing a score scale that associates numbers or other ordered indicators with the performance of examinees. Scaling typically is conducted to aid users in interpreting test results. This module describes different types of raw scores and scale scores, illustrates how to incorporate various sources of information into a score scale, and introduces vertical scaling and its related designs and methodologies as a special type of scaling. After completion of this module, the reader should be able to understand the relationship between various types of raw scores, understand the relationship between raw scores and scale scores, construct a scale with desired properties, evaluate an existing score scale, understand how content and standards information are built into a scale, and understand how vertical scales are developed and used in practice.*

Keywords: raw score, scale score, scaling, vertical scaling

Student A takes the state assessment in year 1 and correctly answers 20 items out of a total of 30 items. Student B takes the state assessment with a different set of test questions (test form) in year 2 and also correctly answers 20 items out of a total of 30 items. By earning the same number-correct raw score on the same state assessment, can we tell whether these two students possess the same level of achievement? We cannot safely make this conclusion because the two test forms from the two years might be somewhat different in difficulty.

When test developers construct test forms, efforts are made to ensure the alternate test forms are as similar as possible from one administration to the next. Even so, alternate test forms generally differ somewhat in difficulty. A number-correct raw score of 20 on one test form does not necessarily indicate the same level of achievement as a number-correct raw score of 20 on another test form. Consequently, information contained in a raw score is limited. Almost all large-scale assessments report *scale scores* to provide information that cannot be reflected in a raw score.

Scaling is the process of constructing a score scale that associates numbers or other ordered indicators with the performance of examinees (Kolen & Brennan, 2004). These numbers and ordered indicators are intended to reflect increasing levels of achievement or proficiency. The process of scaling produces a *score scale*, and the resulting scores are referred to as *scale scores*.

Through the psychometric process of *equating*, the statistical relationship among raw scores on alternate test forms is established. After equating is conducted, a scaling process is applied to convert raw scores on alternate forms to scale scores. The conversion can be a linear or a nonlinear transformation. Test developers can choose which numbers and transformations to use. After the raw scores are converted to scale scores, comparisons can be made from scores earned on alternate test forms. For the example mentioned earlier, Student A's raw score of 20 might be converted to a scale score of 14; Student B's raw score of 20 might be converted to a scale score of 15. This might happen if the alternate form given in year 1 was somewhat easier than the alternate form given in year 2. In this case, we can say that Student B earned a higher scale score than Student A on the test.

Additional information often is incorporated into a scale score during the scaling process. For example, normative information is included so that an examinee's performance can be compared to a relevant reference group. Score precision information is incorporated so that test users will not overinterpret differences among scores. Content information is added so that an examinee's performance can be compared to an established performance standard.

Furthermore, when tests are part of a battery—a set of tests developed together to measure student achievement, such as the Iowa Tests of Basic Skills (ITBS: Hoover, Dunbar, & Frisbie, 2003)—scale scores can allow for interpretation of students' strengths and weaknesses across content areas. Using common scaling conventions for all tests in a battery facilitates computation of *composite scores* across tests—an overall measure of achievement or proficiency.

In some Kindergarten through Grade 12 (K–12) large-scale assessment programs, especially after the implementation of the No Child Left Behind Act of 2001 (NCLB; Public Law 107–110), test developers and users are interested in tracking students' academic growth. *Vertical scales* are constructed to span tests targeting different educational levels. *Vertical scaling* is the process used to construct a vertical scale so that scale scores on tests can be compared across grades. After a meaningful vertical scale is established by the test developer,

Ye Tong, Pearson, 1906 Black Hawk Circle, Audubon, Pennsylvania 19403, United States; ye.tong@pearson.com. Michael J. Kolen, The University of Iowa–Iowa Testing Programs, 224B1 LC, Iowa City, Iowa, 52245, United States, michael-kolen@uiowa.edu.

interpretations such as that a given student grew more from third to fourth grade than he or she did from seventh to eighth grade can be made by test users.

Petersen, Kolen, and Hoover (1989) stated that "the main purpose of scaling is to aid users in interpreting test results" (p. 222). The perspective on scaling taken in this module is consistent with this statement. We take the position that test content and specifications drive test development. Scaling procedures are applied after tests are developed to help facilitate interpretations of the results. In this module, discussions of types of raw scores are presented first followed by the construction of score scales. Vertical scaling designs, methods, and research then are described. In addition, a self-test with answer key is provided.

Raw Scores

Traditionally, the raw score on a test was defined as "the number, proportion, or percentage" of test items that an examinee answers correctly (Petersen et al., 1989, p. 222). In the previous example, both students earned a raw score of 20 on the assessment. Performance assessments, computer-based tests, and the use of item response theory (IRT) (Yen & Fitzpatrick, 2006) have used other types of raw scores.

Raw scores have serious limitations as primary-scale scores for tests. Certain types of raw scores, including number-correct scores, depend on the items that are on a test. When multiple forms of a test exist, such raw scores do not have a consistent meaning across forms. As in the previous example, the two students earning the same number-correct scores on two alternate forms of the same test do not necessarily have the same level of achievement.

When IRT assumptions hold and test items are properly calibrated and placed on the same IRT scale, the IRT proficiency scale does not depend on the items that the student has taken. However, the scaling conventions used with IRT (often unrounded scores with a mean of 0 and a standard deviation of 1) typically do not produce scores that lead to meaningful interpretations without further transformation.

For mixed-format tests, different weights often are applied to scores on the various item types. Test developers can decide on the desired proportional contribution of each item type based on the perceived importance of each item type to the total test, the observed score effective weights, the weights chosen to maximize reliability or other statistical properties, and so on (Kolen, 2006). In the previous example, suppose Student A earned the raw score of 20 by correctly answering 20 multiple-choice (MC) items, and Student B earned the raw score of 20 by correctly answering 16 MC items and one 4-credit constructed-response (CR) item. Depending on the weights that the test developers preassigned to this particular test, the two students could earn quite different scale scores.

When IRT is used with a mixed-format test, the test developer must decide whether the item types are measuring constructs that are similar enough that a unidimensional IRT model can fit to the various item types (Kolen, 2006; Rodriguez, 2003; Wainer & Thissen, 1993). With the exception of the Rasch model, when fitting an IRT model to a mixed-format test, the weighting of each item type can depend on the extent that the item type discriminates near an examinee's proficiency. Test developers need to determine whether these statistical weights reflect the contribution of each item type to the assessment.

Transformations

So far we have discussed various types of raw scores and their related considerations. Different types of transformations are considered in this section. To transform raw scores to scale scores, linear or nonlinear transformations are applied.

Linear Transformations

Linear transformations can be used to transform raw scores to scale scores. One way to linearly transform raw scores to scale scores is to specify scale score equivalents of two raw score points. Defining x_1 and x_2 as these two raw score points and $S(x_1)$ and $S(x_2)$ as the desired scale score equivalents,

$$S(x) = \left[\frac{S(x_2) - S(x_1)}{x_2 - x_1}\right]x + \left\{S(x_1) - \left[\frac{S(x_2) - S(x_1)}{x_2 - x_1}\right]x_1\right\} \quad (1)$$

defines a linear raw-to-scale score equivalent.

Alternatively, if the mean and standard deviation of the scale scores are specified by the test developer and the mean and standard deviation of the raw scores are known, a linear transformation can be established to transform raw scores to scale scores. When it is desirable to specify one scale score equivalent and the standard deviation of the scale scores, the linear transformation can be similarly established. Kolen (2006) provided formulas for these different scenarios.

Nonlinear Transformations

Nonlinear transformations can take on almost any monotonically nondecreasing form, such that the scale score corresponding to a particular raw score is greater than or equal to the scale score corresponding to a lower raw score. One of the least complicated nonlinear transformations is to round the scale scores created by a linear transformation to integers. Another nonlinear transformation is truncation, which is used to ensure that all scale scores are within a desired range.

Rounding and truncation processes are used in many testing programs when the raw-to-scale score transformation begins with a linear transformation. Many testing programs that use IRT-based proficiencies as raw scores linearly transform the IRT proficiency estimates, round these estimates to integers, and truncate the scale at minimum and maximum scores. Equation 1 can be used in such a context, substituting the raw scores with the desired IRT proficiencies. Such a practice often is used with large-scale assessments where cut scores are identified (often two cut scores, Proficient and Advanced) and there are desired scale score values test developers want to use as cut scores.

More complex nonlinear transformations are sometimes used. One such transformation is to convert raw scores to scale scores with a desired distributional shape. One commonly used shape is a normal distribution, and the process of constructing such scores is referred to as the process of normalizing scores. Kolen (2006) provided detailed steps on how such normalized scores are calculated with the desired mean and standard deviation. Some normalized scores include *T*-scores (McCall, 1939), intelligence test scores (Angoff, 1971, pp. 525–526), stanines (Flanagan, 1951, p. 747), and normal curve equivalents.

Information Contained in Score Scales

Incorporating information into score scales is considered in this section. Normative, score precision, and content information can be incorporated to facilitate score interpretation by test users.

Normative Information

The process of incorporating normative information into a score scale begins by administering the test to a group of examinees, commonly referred to as the *norm group*. The norm group might be defined by the test developer as test users at a particular point in time. The norm group chosen for constructing a scale strongly influences the meaning of the resulting scale scores. Linear or nonlinear transformations can be applied, as long as a meaningful norm group has been defined. For example, a scale might be constructed so that the mean of the scale scores is 500, the standard deviation of the scale scores is 100, and the scores are approximately normally distributed.

Score Precision Information

Flanagan (1951) pointed out that scale score units should "be of an order of magnitude most appropriate to express their accuracy of measurement" (p. 746). The choice of the number of scale score units involves using a sufficient number of units to preserve score precision in the raw scores, but not so many that test users attach significance to score differences that are small relative to measurement error.

How many score points should be used? Over the years, various rules of thumb have been developed for choosing the number of scale score units. One of these rules was originally used in developing the scale for the Iowa Tests of Educational Development (ITED, 1958). The ITED scale was constructed in 1942, using integer scores with the property that an approximate 50% confidence interval for true scores could be found by adding and subtracting 1 scale score point to and from an examinee's scale score. Similarly, Truman L. Kelley (W. H. Angoff, personal communication, February, 17, 1987) suggested constructing scale scores so that an approximate confidence interval could be constructed by adding 3 scale score points to and subtracting 3 scale score points from each examinee's scale score. These confidence interval statements were translated into the desired number of discrete scale score points by finding a range of integer scores that are consistent with the confidence interval properties stated. These rules are based on the following assumptions:

1. Linear transformations of raw-to-scale scores are being considered.
2. Measurement error, conditional on true score, is normally distributed.
3. The conditional standard error of measurement (CSEM) is constant along the score scale.
4. The reliability of raw scores, $\rho_{XX'}$, is known or a reasonable estimate exists.
5. The range of scale scores of interest is 6 scale score standard deviations ($6\sigma_S$).
6. The width of the desired confidence interval, symbolized by h, is given. For example, $h = 1$ for the ITED rule and $h = 3$ for Kelley's rule mentioned earlier.
7. The confidence coefficient, γ, is given, and z_γ is the unit-normal score used to form a $100\gamma\%$ confidence interval. For example, $\gamma = 50\%$ and $z_\gamma \approx .6745$ for the ITED rule; $\gamma = 68\%$ and $z_\gamma \approx 1$ for Kelley's rule.

Based on these assumptions, the standard deviation of scale scores is found as

$$\sigma_S = \frac{h}{z_\gamma \sqrt{1 - \rho_{XX'}}}, \tag{2}$$

as shown by Kolen and Brennan (2004, pp. 346–347). Multiplying σ_S by 6 and rounding to an integer gives the number of distinct score points.

For example, assume that $\rho_{XX'} = .91$ and that the ITED rule is being used, where $h = 1$ and $z_\gamma \approx .6745$. Substituting these values into Equation 2 gives $\sigma_S = 4.94$. Multiplying by 6 and rounding to an integer suggests that 30 scale score points should be used given this rule.

Although developed under assumptions of constant and normally distributed measurement errors and linear raw-to-scale score transformations, these rules of thumb have been used under a much wider set of conditions. Keeping in mind that the confidence interval properties are only approximately achieved when the assumptions are not met, these rules are used as approximations when errors are not normally distributed or constant, and when the transformations of raw-to-scale scores are not linear. They can also be used for all types of raw scores, including IRT estimates of θ, as long as a reasonable estimate of reliability is available.

Score scales with approximately equal CSEM. CSEMs index the amount of measurement error involved for the various score points on a test. In general, CSEMs for raw scores are not constant along the score scale (Feldt & Brennan, 1989). For summed scores, the CSEMs tend to be relatively large for middle scores and small for very high and low scores. For IRT-based raw scores, such as maximum likelihood and Bayesian estimates of proficiency, the CSEMs typically follow a pattern opposite to that for summed scores. In this case, the CSEMs are smaller for the middle scores and larger at the extremes. Kolen, Hanson, and Brennan (1992) demonstrated that nonlinear transformations of raw scores can lead to a pattern of CSEMs that is markedly different from that of the raw scores.

In an attempt to simplify score interpretation, Kolen (1988) suggested using an arcsine transformation that stabilizes the magnitude of the CSEM. This transformation leads to CSEMs that are approximately equal along the score scale. If the reporting scale has such properties, then only one overall CSEM needs to be provided to test users instead of multiple values along the score scale.

Content Information

According to Ebel (1962), "to be meaningful any test scores must be related to test content as well as to the scores of other examinees" (p. 18). Ebel recommended that content information be provided along with scale scores to aid score interpretation. Three types of procedures for providing content meaning to scale scores are considered here and referred to as *item mapping*, *scale anchoring*, and *standard setting*.

Item mapping. After constructing a score scale using one of the methods already discussed, items are found that represent

various scale score points. The set of items representing the various scale score points are referred to as item maps. Zwick, Senturk, Wang, and Loomis (2001) reviewed and studied item-mapping procedures.

To implement item maps, test items are associated with various scale score points. The probability of correct response on each item for dichotomously-scored items or each score point for polytomously-scored items is regressed on scale scores using procedures such as logistic regression or IRT. The *response probability* (RP) *level*, which is the probability of correct response given scale score that is associated with mastery on a test item, is stated by the test developer. The same RP level is used for all items of a particular type on the test. Using the regression of item score on scale score, an item is said to map at the scale score that is associated with an RP of correctly answering the item. According to Zwick et al. (2001), values of RP ranging from .50 to .80 have been used for item mapping with the National Assessment of Educational Progress (NAEP). Huynh (1998) provided psychometric justifications for choosing RP. Additional criteria are often used when choosing the items to report on item maps. For example, items might be chosen only if they discriminate well between the examinees who score above and below a particular score. Also, items might be used only if subject matter experts indicate that an item provides adequate representation of test content.

The outcome of an item-mapping procedure is the specification of test questions that represent various scale score points. Item maps were constructed, for example, using an RP level of 74% for MC test questions on the 1996 NAEP fourth-grade Science Assessment (O'Sullivan, Reese, & Mazzeo, 1997).

Scale anchoring. The goal of scale anchoring is to provide general statements of what students who score at each of a selected set of scale score points know and are able to do. The first step in scale anchoring is to create item maps for the items on a test. Then a set of scale score points is chosen. Typically, these points are either equally spaced along the score scale or are selected at a particular set of percentiles. Items that map at or near these points are chosen to represent these points. Subject matter experts review the items that map near each point and develop general statements that represent the skills of examinees scoring at these points. In scale anchoring, it is assumed that examinees know and are able to do all the skills in the statements that are at or below a given score level.

The scale anchoring process used with NAEP is described by Allen, Carlson, and Zelenak (1999). A scale anchoring process was used to develop the ACT Standards for Transition for Explore, PLAN, and the ACT Assessment (ACT, 2001). A process much like scale anchoring was also suggested by Ebel (1962), although he used scores on a subset of items rather than statements to display performance at each level.

Standard setting. Standard-setting procedures begin with *performance-level descriptors*, which are statements of what competent examinees know and are able to do at each performance level. Standard-setting methods seek to find *cut scores*, which are score points that divide the examinees who know and are able to do what is stated from other examinees. Standard setting is often used in professional licensure and certification testing settings. With the implementation of NCLB, standard setting also has become prevalent in K–12 assessments.

Often, a committee of educators or subject experts is assembled, and these judges collectively develop performance-level descriptors; sometimes, as an intermediate step, they develop the most important statements defining students who are barely meeting a certain standard. (There are other variations of this practice.) This exercise is particularly important because it provides a common ground for content experts to make judgments and recommendations on setting the standards. Next, the judges are provided with a set of test questions. A systematic procedure is used to collect information and recommendations from the judges.

In one method, often referred to as the Modified Angoff method, judges are asked to indicate the proportion of the threshold examinees who should be expected to correctly answer each item for a given performance level. The judgments are aggregated over items and judges. The outcome of these procedures typically is a summed score on the set of items that represents the cut point.

The Bookmark procedure (Lewis, Green, Mitzel, Baum, & Patz, 2003) is another prevalent methodology used in standard setting. With this methodology, the set of test questions is ordered from the least difficult question to the most difficult question, and it is often referred to as an ordered item book. An RP value is chosen before the standard-setting meeting. Judges are asked to consider the knowledge and skills the test questions are measuring and their relationship with the performance-level descriptors. The judges then go through the ordered item book to determine, for each test question, whether the threshold students have at least the given RP of correctly answering the item. If the answer is "yes," then the judges move to the next test question on the ordered item book until they reach the item where they consider the answer to be "no." The judges then place their bookmark on the last "yes" item. The bookmark ratings are aggregated across judges for each performance level. The end result is cut scores, typically on an IRT scale.

The judges often go through a total of three rounds of such an exercise before the final cut score recommendations are made. Impact data, in the form of percentage of examinees classified into each performance level based on the recommended cut scores, are provided as a reality check. The cut scores from the last round are considered the final recommendations. Classification consistency and judge variability often are estimated to assess the generalizability of the cut score recommendations. If the standard-setting is conducted for a range of grades, such as grades 3 through 8, then a vertical articulation meeting sometimes follows the standard-setting meetings for each grade. At the vertical articulation session, typically, a subset of committee members from the standard-setting meetings gather to observe the impact of recommended cut scores across the entire range of the grades and determine if the cut scores need to be modified. More recently, with K–12 assessments, some states have chosen to invite policymakers to convene after the standard setting to consider content experts' cut scores together with the impact of their cut scores for a final recommendation.

Standard setting is a judgmental process that incorporates various sources of information, such as content standards, performance-level descriptors, and policy statements. Other popular standard-setting methods include the contrasting groups method and the body of work method (see Cizek,

Bunch, & Koons, 2004, for a summary of standard-setting methods).

Once cut scores are obtained through standard setting, test developers typically set the cut scores to have specific scale score values. A linear transformation (similar to Equation 1) is applied to link raw scores to scale scores. Rounding and truncation are often applied to keep the scores within a certain range. For such a scale, statements about what students should know and are able to do can be made between cut score values, often using performance-level descriptors.

Vertical Scaling

At times, it is desirable to track students' academic growth from one grade to the next. For example, Student A correctly answered 20 items out of 30 items in year 1 and correctly answered 20 items out of 30 items on the same assessment the next year. Does this mean that Student A has made progress? Did Student A demonstrate the same amount of growth as an average student? Again, raw scores do not provide enough information to make such comparisons or interpretations. In this case, a *vertical scale* sometimes is used. Educational achievement and aptitude tests (e.g., ITBS: Hoover et al., 2003; Cognitive Abilities Test: Lohman & Hagen, 2002) typically are constructed using multiple-test levels, where each level is constructed to be appropriate for students at a particular grade or age. *Vertical scaling* procedures are used to relate scores on these multiple-test levels to a common scale.

Vertical scaling is much more complicated than within-grade scaling. Research shows that various decisions in the scaling process tend to lead to different vertical scales with different growth implications (Tong & Kolen, 2007).

Background and NCLB

NCLB requires that all students be tested in reading and mathematics (and other subjects in the future) in grades 3 through 8 and that all states have grade-level content and proficiency standards. Growth modeling is also required under NCLB, with a phase-in plan. NCLB growth modeling does not require a vertical scale. However, due to the desired interpretations, an effective vertical scale can facilitate the growth modeling process. Many states—including Texas, Florida, Minnesota, Colorado, and North Carolina—have considered vertical scales for their grades 3 through 8 assessments.

When a vertical scale is effectively established, it can support the following interpretations:

1. Assessment of individual student growth. Example: Paula's score increased more than John's score from third to fourth grade.
2. Mapping of items and skills within and across grades on a single scale. Example: The fourth-grade items covering content domain A are more difficult than sixth-grade items covering content domain B.
3. Relating proficiency standards from different grades to a single scale. Example: A student who is "advanced" in fourth grade is higher achieving than a student who is "proficient" in fifth grade.

Despite the difficulty of establishing a vertical scale, test developers continue to explore ways to construct effective and psychometrically sound vertical scales because of their many desired properties.

Definition of Growth

There is not a single prevalent definition for students' academic growth. The nature of growth is more of a theoretical issue than a measurement issue and is affected by such considerations as child development, psychology, and how school curricula are implemented (Harris, Hendrickson, Tong, Shin, & Shyu, 2004).

Kolen and Brennan (2004) discussed two types of definitions for growth. Considering a domain of content covered by an achievement test over all grade levels, the *domain definition of growth* refers to growth over all the content in the domain. Conversely, the *grade-to-grade definition of growth* is defined over the content that is on a test level appropriate for students in a particular grade. Therefore, under the domain definition, growth by a student from one year to the next is defined over the entire content domain across all grades. Under the grade-to-grade definition, growth is assessed using the content included in adjacent grades. For subject matter areas that are closely related to the school curriculum, growth observed between adjacent grades tends to be different under the two definitions of growth.

Data Collection Designs

A special data collection design usually is involved when a vertical scale is being developed. In this module, three common designs for data collections are described and compared.

Common-item design. Following this design, each test level is administered to students at the appropriate grade. Common items between adjacent levels are included on the test. Performance on the common items is used to indicate the average amount of growth that occurs from one grade to the next. One of the key decisions with this design is the selection of common items: should common items be selected from both grades, or should the common items be selected from only one grade? Opponents of the first approach argue that students from lower grades should not be tested on items for which they have not received instruction; opponents of the second approach argue that such a common item set may not achieve good representation of the content. Additionally, decisions need to be made on whether the common items should be selected as a miniature of the entire test, as typically seen in an equating setting, or to be a good representation of the domain overlap between the two adjacent grades. As described in Tong and Kolen (2007), different decisions are likely to lead to vertical scales with different properties.

To implement the common-item design, one grade level is designated as the base level. Students' performance on the common items is used to link scores from adjacent grades. A chaining process is then used to link scores from all levels to scores on the base level. For example, if grade 5 was selected to be the base level, to link grade 7 scores to the base scale, grade 7 scores are placed onto the same scale as grade 6, through items that are common between grades 6 and 7 tests. Next, grade 6 scores are placed onto the grade 5 score scale through the items that are common between grades 5 and 6. Using this chain, scores on grade 7 are placed onto the grade 5 base scale. Many of the vertical scales developed in K–12 assessments use the common-item design.

Equivalent groups design. In the *equivalent groups design*, examinees in each grade are randomly assigned to take either the test level designed for their grade or the test level designed for adjacent grades. Through the random equivalence of the groups, students' performance on the tests from adjacent levels can help establish the average growth from one grade to the next. A base grade is again chosen, and by chaining across grades, the data from this administration are used to place scores from all test levels onto the same scale. Note that this design does not necessarily require common items between adjacent grades. Instead, it uses common people to establish the link. Because this design requires students to be administered a test level not designed for their grade, it is not used very often in practice.

Scaling test design. In the *scaling test design*, a special test called a *scaling test* is constructed that spans the content across all grade levels of interest. Students in all grades are administered the same scaling test. In addition, students also take a test that is appropriate for their grade. The score scale is defined using scores on the scaling test. The level-specific tests serve to better measure students' proficiency. The challenge when using this design is to construct the special scaling test that covers the entire content domain from lower to higher grades. The scaling test design is used to construct the vertical scale for the ITBS.

Similarities and differences. The common-item design can be implemented within an operational setting. The equivalent groups design requires a special administration. The scaling test design requires construction of a scaling test and a special administration. For both the scaling test design and the equivalent groups design, students can take many test questions that are not appropriate for the grades they are in.

Although it is the most difficult to implement, the scaling test design has the advantage of explicitly considering the domain definition of growth by ordering students from all grades on a single test. The other two designs do not allow for an explicit ordering because examinees in different grades take different test questions. The scaling test design is expected to produce scaling results that are different from those produced by the other two designs, especially for content areas that are closely tied to the curriculum. The growth implications may also differ across the designs (Tong & Kolen, 2007).

Scaling Methods

After the test is constructed and data collected, statistical methods are used to construct the vertical scale. In this section, three general statistical approaches are considered. Within each of these general statistical approaches, the specific procedures that are used depend on the data collection design.

Hieronymus scaling. The developmental score scale used on the ITBS (Hoover et al., 2003) was constructed using the Hieronymus scaling method that was originally developed by Albert Hieronymus (Petersen et al., 1989, p. 232). Under this scaling method, test developers start with a model that best reflects the nature of the assessment. Although it can be used with any data collection design, this method was developed specifically to be used with the scaling test design. To conduct the scaling, the median summed score on the scaling test for each grade level is assigned a prespecified scale score value. These prespecified values are based on the test developer's expectations on how year-to-year median growth changes based on the construct being measured. For example, with the current forms of the ITBS, the test developer's model for growth is that the average amount of growth from year-to-year decreases over grades and that the students become more variable in academic achievement. A vertical scale is built to reflect such a model, with decelerating growth and increasing within-grade standard deviations.

After the conversion of scaling test scores to scale scores is found, scores on each of the level tests are related to the vertical scale through the scaling test. As described by Petersen et al. (1989), Hieronymus scaling uses distributions of Kelley regressed score estimates of true scores on the scaling test and the level tests in this linking process. Because the level tests are targeted to the examinees at each grade, they tend to yield scores at each grade that are more reliable than the scores on the scaling test. Using estimated true scores is intended to handle these differences in reliability.

Thurstone scaling. Thurstone (1938) described a scaling method that involves normalizing summed scores within each grade. Using any of the three data collection designs, this method is based on the assumption that scale scores are normally distributed within grade and are linearly related across grades. Summed scores are first converted to normalized z scores. A subset of z scores is used in the Thurstone scaling. Gulliksen (1950, p. 284) recommended choosing 10 or 20 raw score points when implementing this procedure and constructing scatter plots for the 10 or 20 z score pairs between adjacent grades to determine if the relationship is linear. Williams, Pommerich, and Thissen (1998) found that the points chosen can affect the characteristics of the resulting scale.

IRT scaling. IRT scaling makes use of the entire set of item-level responses from examinees to the test questions. It can be applied using any of the three data collection designs. For any of the designs, various procedures can be used to estimate IRT item parameter and examinee proficiency (Hambleton & Swaminathan, 1985; Lord, 1980; van der Linden & Hambleton, 1997).

Two calibration approaches are often used for item parameter estimation. In *separate calibration*, item parameters are estimated separately for each grade level. A chaining process is applied to place all grade levels onto the same scale. In *concurrent calibration*, data for examinees across all grades are calibrated in one computer run. Grade groups are identified in the data so that the program allows for separate proficiency distributions for each grade. After concurrent calibration, all item parameters are on the same IRT scale, and no further transformation is needed.

Kolen and Brennan (2004) pointed out that concurrent calibration is often easier to implement, uses the maximum amount of information, and produces more stable linking results when the IRT model holds (Hanson & Béguin, 2002; Kim & Cohen, 1998). However, when the IRT model does not hold, Béguin, Hanson, and Glas (2000), Béguin and Hanson (2001), and Tong and Kolen (2007) found that separate calibration was more accurate. Kolen and Brennan (2004, p. 391) suggested that separate calibration might be

preferable because it mitigates the effects of violations of the unidimensionality assumption in IRT. Furthermore, separate calibration allows for comparison of different item parameter estimates for the common items, which often is used to identify items that are behaving differently in adjacent grades. In addition, concurrent calibration often has convergence problems, especially when conducted across many grades.

Various approaches exist for estimating examinee proficiency. One decision made when applying many of the IRT methods is whether to use raw scores as a function of summed scores (summed scoring) or a function of the whole response pattern (pattern scoring). A decision is also made about whether to use maximum likelihood, Bayesian, or some other type of proficiency estimate. The summed scoring approach produces a one-to-one relationship between raw scores and scale scores and is much easier to explain to test users and the public. Pattern scoring uses more information and tends to be more accurate, but is less easily explained to users.

Tong and Kolen (2007) empirically compared scoring procedures and proficiency estimators with data from an achievement test battery and simulated data. With respect to growth patterns, their results showed little difference between summed scoring and pattern scoring. Nontrivial differences, however, were detected among proficiency estimators, especially between the Maximum Likelihood Estimator (MLE), $\hat{\theta}_{MLE}$, and the Expected A Posteriori (EAP) estimator, $\hat{\theta}_{EAP}$. Both $\hat{\theta}_{MLE}$ and $\hat{\theta}_{EAP}$ depend on the entire item response string; $\hat{\theta}_{MLE}$ does not exist for 0 or perfect scores; $\hat{\theta}_{EAP}$ is a biased Bayes estimator, regresses toward the mean, and tends to be affected by the prior distributions. Based on how these estimators function, Kolen and Tong (2010) pointed out the following patterns in marginal variance hold if the distribution of proficiency is well specified:

$$\text{var}(\hat{\theta}_{TCF}) > \text{var}(\hat{\theta}_{MLE}) > \text{var}(\hat{\theta}_{EAP}) > \text{var}(\hat{\theta}_{sEAP}). \quad (3)$$

In this equation, $\hat{\theta}_{TCF}$ refers to the proficiency estimate based on the test characteristic function and is a summed scoring estimator. $\hat{\theta}_{sEAP}$ refers to the proficiency estimate based on the summed scoring using a Bayesian approach (Thissen & Orlando, 2001). These characteristics of the estimators have implications for summary statistics for the resulting vertical scales.

Under the common-item design, test levels either can be separately calibrated and linked using common items or calibrated concurrently. Under the scaling test design, there are a few possible approaches that include the following: (a) concurrently calibrate the scaling test across grades and separately calibrate the level tests for each grade, using the scaling test results to establish the link; (b) concurrently calibrate the scaling test and level tests; or (c) separately calibrate the scaling test for each grade, separately calibrate level tests, and use the scaling test results to establish the link.

Scaling methods comparisons. All the scaling methods can be used with any of the data collection designs. They all make certain assumptions about score distributions, the underlying construct, and item responses, and they all use some type of statistical technique to establish the link among the grades. Hieronymus scaling has only been used to scale the ITBS test battery; Thurstone scaling has not been used very often in recent years; and most of the existing vertical scales have been established through an IRT scaling method. IRT scaling involves decisions at multiple stages of the scaling process, such as choosing an IRT model, a calibration approach, and a scoring approach. The IRT method is also based on stringent item-level assumptions.

Different scaling methods can produce vertical scales with different characteristics. Growth patterns and within-grade variability are two main features of vertical scales. Hieronymus scaling defines the characteristics of the scale ahead of time and can produce a scale with desired properties that test developers consider appropriate for the assessment. Thurstone scales tend to produce decelerating growth patterns and increasing within-grade variability (Tong & Kolen, 2007; Yen 1986; Yen & Burket, 1997). IRT methods tend to produce scales with decelerating growth (Tong & Kolen, 2007; Williams et al., 1998). In terms of within-grade variability, IRT scaling produces scales with an irregular trend (Becker & Forsyth, 1992; Bock, 1983; Camilli, Yamamoto, & Wang, 1993; Seltzer, Frank, & Bryk, 1994; Williams et al., 1998; Yen & Burket, 1997) or a decreasing trend, referred to as scale shrinkage (Andrews, 1995; Hoover, 1984; Tong & Kolen, 2006, 2007; Yen, 1986).

Kolen and Brennan (2004) concluded that "research suggests that vertical scaling is a very complex process that is affected by many factors. These factors likely interact with one another to produce characteristics of a particular scale. The research record provides little guidance as to what methods and procedures work best" (p. 418). The choice of scale should also be based, at least in part, on educational theory about the construct measured by the test.

Concluding Comments

Yen (1986) pointed out that "choosing the right scale is not an option. It is important that any choice of scale be made consciously and that the reasons for the choice be carefully considered. In making such choices, appealing to common sense is no guarantee of unanimity of opinion or of reaching a sensible conclusion" (p. 314). The perspective taken in this module is that the purpose of scaling is to aid users in interpreting test results. The incorporation of norms, score precision, and content information into score scales are used to help achieve this purpose.

Self-Test

1. A state has two assessment programs. One has a scale based on state norms, with the mean set at 50 and the standard deviation set at 15. The other program has a scale based on cut scores from a standard setting, with "proficient" corresponding to a scale score of 50 and "advanced" corresponding to a scale score of 70. After the scale for the second program is developed, the standard deviation of the score distribution is also 15. Kelly took the first assessment and obtained a scale score of 65; David took the second assessment and also obtained a scale score of 65. Can we say that Kelly and David demonstrated the same level of achievement on these two tests? What interpretations can be made regarding their scale scores?
2. A state has an assessment program from grades 3 through 8. Standard setting was conducted separately for each grade,

with some type of vertical articulation across grades at the end. The proficiency cut score was set to have a scale score value of 300 for grade 3, 400 for grade 4, 500 for grade 5, 600 for grade 6, 700 for grade 7, and 800 for grade 8. Lily scored 300 when she was in grade 3, and she scored 500 when she was in grade 4. Is it reasonable to assume that Lily has reached the proficiency level for grade 5 based on her test score from grade 4?

3. For a test with a reliability of .70, what would be the appropriate number of score points for a scale where a 90% confidence interval is to be constructed by adding and subtracting 2 scale score points?
4. A test has raw scores ranging from 0 to 35.
   a. The test developer would like to use a linear transformation to obtain a scale with the lowest scale score set at 100 and the highest scale score set at 300. If such a scale is developed, what is the scale score value corresponding to a raw score of 20?
   b. After standard setting, the raw score of 18 was determined to be the cut score for "proficient," and the raw score of 25 was determined to be the cut score for "advanced." The test developer would like to apply a linear function to transform raw scores to scale scores, and they would also like to set "proficient" to have a scale score value of 200, and "advanced" to have a scale score value of 250. Once the scale is developed, what is the scale score value corresponding to a raw score of 20?
   c. What different interpretations can be made regarding a raw score of 20 based on the two different scaling processes?
5. A vertical scale was established for a state test from grades 3 through 8. The test developer must decide whether to use summed scoring or pattern scoring and which IRT estimator to use. If you are the psychometrician making recommendations, what characteristics of the following estimators would you highlight to the test developers?
   a. test characteristic function,
   b. MLE,
   c. EAP based on pattern scoring.

Answers to Self-Test

1. Without further information, it is difficult to determine whether the two assessments are based on the same construct, test specifications, and standards. Just because Kelly and David obtained the same scale scores on two separate assessments, this does not provide information regarding their achievement levels in comparison to one other. With Kelly's score, a valid interpretation is that Kelly scored 1 standard deviation above the mean. If the score distribution is normal, Kelly's percentile rank on this assessment should be around 84. David's score tells us that he has mastered the knowledge and skills related to the "proficiency" performance level on the assessment, but that he has not quite yet reached the "advanced" requirement.
2. We cannot make that judgment. The assessment program does not have a vertical scale. Vertical articulation is not vertical scaling. An appropriate interpretation that we can make is that Lily was right at the proficiency level when she was in grade 3, but that she had surpassed the proficiency level—possibly by quite a bit, depending on the characteristics of the score distribution in grade 4 and how large the standard deviation is—when she was in grade 4.
3. $\sigma_S = \frac{h}{z_\gamma \sqrt{1-\rho_{XX'}}} = \frac{2}{1.64\sqrt{1-.70}} = 2.2$. Therefore, $6(2.2) = 13.2$, approximately 13 score points.
4. a. Using Equation 1,

$$S(x) = \left[\frac{S(x_2) - S(x_1)}{x_2 - x_1}\right]x + \left\{S(x_1) - \left[\frac{S(x_2) - S(x_1)}{x_2 - x_1}\right]x_1\right\}$$
$$= \frac{300 - 100}{35 - 0}x + 100 = 5.71x + 100.$$

   Therefore, a raw score of 20 corresponds to a scale score of $S(x) = 5.71 * 20 + 100 \approx 214$.

   b. Again, using Equation 1,

$$S(x) = \left[\frac{S(x_2) - S(x_1)}{x_2 - x_1}\right]x + \left\{S(x_1) - \left[\frac{S(x_2) - S(x_1)}{x_2 - x_1}\right]x_1\right\}$$
$$= \frac{250 - 200}{25 - 18}x + \left\{200 - \left[\frac{250 - 200}{25 - 18}\right]18\right\}$$
$$= 7.14x + 71.43.$$

   Therefore, rounding to an integer, a raw score of 20 corresponds to a scale score of $S(x) = 7.14 * 20 + 71.43 \approx 214$.

   c. For part *a* of the question, no interpretation can be attached to the scale score of 214 obtained. For part *b*, we can say that if a student scores 214, this student has mastered the knowledge and skills related to the proficiency cut score of 200 and should know and be able to do the requirements based on the performance-level descriptors for proficiency. This student has not yet reached the advanced performance level.
5. Different estimators have different statistical properties. Before adoption of a certain IRT estimator, its characteristics, and its effect on the resulting score distributions should be examined and discussed.
   a. The test characteristic function estimator, $\hat{\theta}_{TCF}$, is a summed scoring function. It is monotonically related to the summed score. There is a one-to-one relationship between raw score and the *TCF* estimate, which makes interpretation relatively straightforward for test users. This estimator likely contains more estimation error than pattern-scoring-based estimators, and it tends to produce relatively large variability of score distributions compared with other estimators.
   b. The MLE, $\hat{\theta}_{MLE}$, depends on the entire pattern of item responses. It might be difficult to explain to test users why the same number-correct raw score does not necessarily translate into the same scale score. In addition, no estimates exist for a score of 0 or a perfect score, so an adjustment is made for the two extreme scores. This estimator tends to produce a score distribution with slightly smaller variability than the score distribution produced based on *TCF*.

c. The EAP estimator, $\hat{\theta}_{EAP}$, is based on the pattern of the item responses. Similar to $\hat{\theta}_{MLE}$, the same raw score does not typically translate to the same scale score if they are based on different item response patterns. As a Bayes estimator, *EAP* is a biased estimator if the test is less than perfectly reliable but tends to have smaller root mean square error. It is also a shrinkage estimator and tends to be influenced by the prior distribution used to conduct the estimation. This estimator also tends to produce a score distribution with less variability than does $\hat{\theta}_{MLE}$.

References

ACT (2001). *EXPLORE technical manual*. Iowa City, IA: Author.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report*. Washington, DC: National Center for Education Statistics.

Andrews, K. M. (1995). *The effects of scaling design and scaling method on the primary score scale associated with a multi-level achievement test*. Unpublished Ph.D. dissertation, The University of Iowa.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, *29*(4), 341–354.

Béguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the Meeting of the American Educational Research Association, New Orleans, LA.

Bock, R. D. (1983). The mental growth curve reexamined. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 205–209). New York: Academic Press.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*(4), 379–388.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50.

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement*, *22*(1), 15–25.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24.

Harris, D. J., Hendrickson, A. B., Tong, Y., Shin, S., & Shyu, C. (2004, April). *Vertical scales and the measurement of growth*. Paper presented at the Annual Conference of the National Council of Measurement in Education, San Diego.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, *3*(4), 8–14.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa Tests: Guide to research and development*. Chicago, IL: Riverside Publishing.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational & Behavioral Statistics*, *23*(1), 35–56.

Iowa Tests of Educational Development (1958). *Manual for school administrators. 1958 revision*. Iowa City, IA: The University of Iowa.

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131–143.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*(2), 97–110.

Kolen, M. J. (2006). Scales and norms. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education/Praeger.

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

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*(4), 285–307.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*(3), 8–14.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (2003, April). *The Bookmark procedure: Methodology and recent implementations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Lohman, D. F., & Hagen, E. P. (2002). *Cognitive abilities test. Form 6. Research handbook*. Itasca, IL: Riverside Publishing.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

McCall, W. A. (1939). *Measurement*. New York: Macmillan.

No Child Left Behind Act of 2001, Public Law No. 107-110, 115 Stat. 1425 (2002).

O'Sullivan, C. Y., Reese, C. M., & Mazzeo, J. (1997). *NAEP 1996 science report card for the Nation and the States*. Washington, DC: National Center for Education Statistics.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163–184.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation Policy Analysis*, *16*(1), 41–49.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Erlbaum.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Tong, Y., & Kolen, M. J. (2006, April). *Scale shrinkage*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Francisco.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*(2), 227–253.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, *35*(2), 93–107.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299–325.

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement*, *34*(4), 293–313.

Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15–25.