Chapter 4
Statistical Models for Vertical Linking

James E. Carlson

4.1 Introduction

Vertical linking, sometimes referred to as vertical scaling or cross-grade scaling, comprises a variety of techniques used to develop and maintain vertical scales that are developmental in nature, encompassing two or more grades in schools. Separate tests designed to measure achievement on the same dimension at each grade level are linked through various procedures to enable the measurement of growth across the levels. Formal equating, having the goal of interchangeability of scores on different test forms, is not possible for vertical linking because interchangeability is not feasible in this context: The appropriate content of the tests for the different grade levels necessarily differ because the curricula differ. In addition, the difficulty levels of tests at two adjacent grade levels are typically different, so the tests cannot be parallel as required for a formal equating. Most of the designs used to accomplish vertical linking do, however, involve some content that is appropriate for adjacent grade levels.

Several of the statistical procedures discussed in other chapters of this work, for example item response theory (IRT), can be applied to the problem of vertical linking. Although non-IRT approaches such as equipercentile methods can be used, most vertical linking is done in large-scale assessment programs that use IRT scaling, so those methods will be the focus of this chapter.

Although grade-level tests are used in discussions here, note that several test publishers have developed vertical scales comprising levels each of which may be administered at several grade levels. The designs for vertical linking discussed in this chapter all use cross-sectional data. That is, the data are assumed to be collected during a given time period using independent samples from the different grade levels. An alternative that has not, to my knowledge, been used is a longitudinal

J.E. Carlson
Educational Testing Service, Rosedale Rd, Princeton, New Jersey 08541, USA
e-mail: jcarlson@ets.org

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_4,

design in which data are collected in successive years from the same cohort of students. For example, the same student cohort is tested at the third grade in one year and the fourth grade in the following year. With such a design, the samples will not be independent and data will not be available on all students for both years, due to students' transferring schools or being held at the same grade level. Such characteristics of longitudinal samples must be taken into consideration when using those designs.

4.2 Designs for Vertical Linking

Holland and Dorans (2006, pp. 197-201), Kolen (2006, pp. 173–174), and Kolen and Brennan, 2004, pp. 372–418) described several common-item and equivalent-groups designs that may be used for collecting the data necessary to develop a vertical scale. There are alternatives to those discussed in these sources. One example of the common-item design and one of the equivalent-groups design will be used as illustrations.

In both designs, a linking procedure typically involves starting by linking scales in two adjacent grades (e.g., the second- to the first-grade level), continues by linking scales in the next grade (e.g., third grade to second), and proceeds similarly until scales in all grades in the assessment are linked to form the vertical scale. The linking can begin at any grade level, for example starting with a middle grade and linking upward and downward from there. Another design discussed by Kolen and Brennan (2004), the scaling test design, involves administering one test form (the scaling test) to all grades. The results of that administration are used to set the vertical scale. The scales of test forms at each grade level are subsequently linked to that vertical scale. In my opinion, this design is not appropriate for most educational assessments because it involves testing most students with a number of items that are too difficult or too easy for them, hence yielding no information for those students. An additional issue involves testing students on content that they have not had the opportunity to learn. Although this is partially an ethical issue, the lack of information yielded by the data is also a technical issue that is mentioned where relevant to each design.

4.2.1 Common-Item Designs

In a common-item design, a group of students at each grade level is selected to be administered blocks of items. Some blocks comprising the common items (referred to as anchor blocks) are administered at adjacent grade levels. One form of this design is illustrated in Table 4.1. In this design students at each grade level are administered some blocks of items unique to their grade level and some

Table 4.1 A Common-Item Design With On-Grade and Anchor Item Blocks

Student grade	On-grade item block	Anchor (linking) item block					
		G3-4	G4-5	G5-6	G6-7	G7-8	G8-9
3	G3	X	–	–	–	–	
4	G4	X	X				
5	G5		X	X			
6	G6			X	X		
7	G7				X	X	
8	G8					X	X
9	G9						X

anchor-item blocks shared with adjacent grade levels. The latter provide the vehicle for linking.

For example, third-grade students take a block (G3) of third-grade items and a linking block (G3-4) of items appropriate for Grades 3 and 4. The fourth-grade students take a block (G4) of fourth-grade items and two linking blocks: one of items appropriate for Grades 3 and 4, and the other of items appropriate for Grades 4 and 5 (G4-5). Using scores on the linking block (G3-4) from students from Grades 3 and 4, scales from G3 and G4 can be linked. Saying the G3-4 block is appropriate for Grades 3 and 4 means that students at both grade levels have had the opportunity to learn the content being tested.

4.2.2 Equivalent-Groups Design

In the equivalent-groups design, randomly equivalent groups of students at the same grade level are administered different blocks of items, and most of the blocks are administered to groups at adjacent grade levels. One example of this design is illustrated in Table 4.2. In this design one sample of students at each grade level takes a test form including a block of items in common with the adjacent grade below, one sample takes item blocks only for that grade, and the third sample takes a block of items in common with the adjacent grade above. For the lowest and highest grades in the design they can, of course, only share item blocks with one adjacent grade. For example, the shaded portion of Table 4.2 shows that there are three randomly equivalent samples at the fourth grade. Sample 4a takes a block (3B) of third-grade items and a block (4A) of fourth-grade items, Sample 4b takes two blocks (4A, 4B) of fourth-grade items, and Sample 4c takes a block (4B) of fourth-grade items and a block (5A) of fifth-grade items. Two blocks within each grade are used for illustrative purposes; different assessments will use different numbers of blocks depending on issues such as content coverage and the need for different forms because of security issues. One important aspect of the content coverage issue for vertical scaling designs is that administering items at a grade level above that of the students would not be appropriate if the students have not

Table 4.2 Equivalent Groups Design with Common Blocks of Items at Adjacent Grades

Student grade	Student samples	Item blocks by grade													
		G3		G4		G5		G6		G7		G8		G9	
		3A	3B	4A	4B	5A	5B	6A	6B	7A	7B	8A	8B	9A	9B
3	3a	X	X												
	3b		X	X											
4	4a		X	X											
	4b			X	X										
	4c				X	X									
5	5a				X	X									
	5b					X	X								
	5c						X	X							
6	6a						X	X							
	6b							X	X						
	6c								X	X					
7	7a								X	X					
	7b									X	X				
	7c										X	X			
8	8a										X	X			
	8b											X	X		
	8c												X	X	
9	9a												X	X	
	9b													X	X

Note. Student samples randomly equivalent within grade.

been exposed to the relevant content in the item blocks. To do so would not yield any useful information for scaling or for scoring. Hence, for example, Block 4A in Table 4.2 must contain content appropriate for both third- and fourth-grade students. In order to cover all fourth-grade content in the assessment, however, typically additional blocks at that grade level must cover content not appropriate for administration to students in Grades 3 or 5.

Note that Kolen and Brennan (2004) illustrated a simpler equivalent-groups design that has common items only with the grade below, so the lowest grade shares no blocks with an adjacent grade. One problem with their design is that all item blocks except one each for the lowest and highest grades must be comprised of content appropriate for two grade levels. That is, the design does not allow for items with content appropriate for a single grade level. As discussed above, adequately covering all the important curricular content at each grade level likely requires including blocks of items testing content that is only appropriate for one of the grades in the assessment. Other variations on this design are also possible.

4.3 IRT Models for Vertical Linking

Methods based on IRT models discussed in other chapters of this work are most commonly used in vertical linking (see also Holland & Dorans, 2006; Kolen, 2006; Patz & Yao, 2007; Thissen & Steinberg, 1986; Yen & Fitzpatrick, 2006). Although

those sources provide discussion of many models, this chapter focuses on the models most commonly used in operational educational assessment programs. Either the common-item or the equivalent-groups design may be used to gather the data, as described below.

Most of the IRT models commonly used in assessments using vertical scaling are special cases of a model that can be written in one general form. Define

$$\begin{aligned} f_{jk} &= Da_j(\theta - b_{jk}) \\ &(k = 0,\ 1,\ 2,\ ...,\ m_j - 1) \\ b_{j0} &= 0.0 \end{aligned} \tag{4.1}$$

where D is a scaling factor of 1.7 (Haley, as cited in Lord & Novick, 1968, p. 399; specified so that the logistic and normal ogive models differ by less than .01 for all θ values), a_j is the discrimination parameter for item j, θ is the proficiency variable, and b_{jk} is a location parameter for the kth-level of item j having m_j score levels numbered from zero to m_j-1. Then, the general form of the logistic model (an alternative is a similar normal model; see Lord & Novick, 1968) for item j is

$$\begin{aligned} P_{jk}(\theta) &= c_j + \frac{\left(1 - c_j\right) e^{\sum_{t=0}^{k} f_{jt}}}{\sum_{s=0}^{m_j-1} e^{\sum_{t=0}^{s} f_{jt}}} = c_j + \frac{\left(1 - c_j\right) e^{f_{j0}}\, e^{f_{j1}} \ldots e^{f_{jk}}}{e^{f_{j0}} + e^{f_{j0}} e^{f_{j1}} + \ldots + e^{f_{j0}} e^{f_{j1}} \ldots e^{f_{jm_j-1}}} \\ &= c_j + \frac{\left(1 - c_j\right) \prod_{t=0}^{k} e^{f_{jt}}}{\sum_{s=0}^{m_j-1} \prod_{t=0}^{s} e^{f_{jt}}} \end{aligned} \tag{4.2}$$

where c_j is the lower asymptote parameter. Note that for all two-parameter models c_j is zero, including two equivalent models that were independently developed at about the same time: Yen's (as cited in Yen & Fitzpatrick, 2006) two-parameter partial-credit model and Muraki's (1992) generalized partial-credit model.[1] Yen's model defines the expression in Equation 4.1 as

$$\begin{aligned} f_{jk} &= k\alpha_j\theta - \sum_{t=0}^{m_j-1} \gamma_{jt} \\ &(k = 1,\ 2,\ 3,\ ...,\ m_j - 1) \\ \gamma_{j0} &= 0.0, \end{aligned}$$

[1]Yen developed her model in 1991 (published in a technical report in 1992, as cited in Yen & Fitzpatrick, 2006).

whereas Muraki defines it as

$$f_{jk} = a_j(\theta - b_j + d_{jk})$$
$$d_{j0} = 0.0,$$

The three parameterizations of the model are related as follows:

$$\alpha_j = a_j$$
$$\gamma_{jk} = a_j b_{jk}$$
$$b_j = \frac{1}{m_j - 1} \sum_{k=1}^{m_j - 1} b_{jk}$$
$$d_{jk} = b_j - b_{jk} \quad (k = 1, 2, 3, ..., m_j - 1),$$

Note also that for a dichotomously scored item m_j is 2 and Equation 4.2 reduces to one of the logistic models: one-parameter logistic (1PL), 2PL, and 3PL. With a, b, and c all present it is the 3PL; with c set to zero it is the 2PL; and if a is set to 1.0 (a actually can be any constant value across all items) and c to 0.0, it is the 1PL.

Most, if not all, of the IRT models discussed by Thissen and Steinberg (1986) and Yen and Fitzpatrick (2006, pp. 113-118), and in other chapters of this book, can be used to fit an IRT model to the data used in vertical scaling. A competitor to the two-parameteror generalized partial-credit model is Samejima's (1969) graded-response model, which is used for vertical scaling purposes in a number of educational assessment programs. Again, the graded-response model can be expressed in either normal ogive or logistic form, and the latter can be expressed as

$$P_{jk}(\theta) = \left[1 + e^{-f_{jk}}\right]^{-1} - \left[1 + e^{-f_{j,k+1}}\right]^{-1},$$

where f_{jk} and $f_{j,k+1}$ are as defined in Equation 4.1.

4.3.1 The Common-Item Design

As mentioned above, in the common-item design linking is carried out through an anchor block of items administered at adjacent grade levels under the assumption that the parameters of the items are common to the two levels. Also as mentioned above, the content covered in the anchor block of items must be appropriate for students at both grade levels to avoid testing some students with items that yield no information, due to lack of opportunity to learn in their grade-level curriculum. Fitting the IRT model to the data is usually referred to as a calibration of the items and results in estimates of parameters for each item. The item parameter estimates for the anchor block of items in two adjacent grade levels are then used to perform

the linking. When underlying assumptions are satisfied and the tests of two adjacent grade levels are separately calibrated, as described by Kolen (2006, p. 176), the parameters of the anchor items at the two levels are linearly related and therefore can be placed on the same scale via a linear transformation. The item parameters and proficiency variable are transformed via

$$\theta^* = K_2 + K_1\theta$$
$$b^* = K_2 + K_1 b$$
$$a^* = \frac{a}{K_1},$$

where the quantities with the asterisks represent the transformed quantities and the constants, K_1 and K_2 are determined by the specific linear transformation method employed. Methods of doing this include mean-mean, mean-sigma, and the Stocking-Lord and Haebara test characteristic-curve (TCC) methods (see Yen & Fitzpatrick, 2006, pp. 134-135; Kolen & Brennan, 2004, pp. 387-388).

An alternative to the linear transformation methods is concurrent calibration, in which data from several grades are calibrated together. This method assumes that the anchor items have the same parameters in each grade to which they are administered.

To link all grades via linear transformation methods, the process begins by defining one grade level as the base level and proceeding to link the other grades in a chain of transformations. For example, to link Grades 3–9, the Grade 3 calibration results could be used to define the base scale. Using the anchor items common to Grades 3 and 4, the linear transformation method would be used to transform the anchor items' calibration results in the fourth grade to the third-grade scale. Nonanchor items on the fourth-grade test would undergo the same transformation to place them on the scale. A similar procedure would be used to place the fifth-grade results on the scale, using the anchor items common to Grades 4 and 5. This procedure would be continued until the scale encompassed all seven grade levels.

4.3.1.1 Mean-Mean and Mean-Sigma Methods

As mentioned by Kolen and Brennan (2004), the mean-mean and mean-sigma methods use the means, or means and standard deviations, respectively, of the location parameter estimates for the anchor items to define the linear transformations. The transformation constants are defined as

$$K_2 = \frac{s_t}{s_u}$$
$$K_1 = M_t - K_2 M_u,$$

where N_t and s_t represent the mean and standard deviation of the target location parameters and M_u and s_u represent the mean and standard deviation of the untransformed location parameter estimates. For the mean-mean method, K_2 is set to 1.0 and K_1 is simply the difference between the target mean and the untransformed mean of the location parameters.[2]

As mentioned by Kolen and Brennan (2004, p. 168) and Yen and Fitzpatrick (2006, pp. 134-145) these simple linear transformations of parameter estimates can be problematic in that different combinations of these IRT parameter estimates can result in very similar item response functions (IRFs, discussed below). The TCC methods avoid this problem by using the IRFs and TCCs rather than the individual parameter estimates. Yen and Fitzpatrick also mentioned that the TCC methods have the advantage of "using weights for the minimization based on a distribution of abilities, so that more weight is given in parts of the scale where there are more examinees" and minimize "differences in expected scores rather than observed scores or parameters" (p. 135). I prefer the TCC methods for another reason mentioned by Yen and Fitzpatrick. In many assessments, the TCC is the basis for estimates of examinees' scores using estimation methods presented by Yen (1984; see also Yen & Fitzpatrick, 2006, p. 137). Hence, matching the TCCs for the anchor items on two forms of a test provides a criterion that is directly related to commonly used scoring procedures. To apply the TCC methods, separately calibrate the data within samples of the two datasets whose scales are to be linked. The TCC in one group serves as the target for the transformation, and the other is to be transformed to match as closely as possible that target TCC. The untransformed and transformed TCCs in the latter group are usually referred to as the *provisional* and *transformed curves*.

4.3.1.2 The Stocking-Lord TCC Method

The TCC method of Stocking and Lord (1983) is probably the most widely used method for vertical linking through anchor items. One formulation of the criterion for the Stocking-Lord method is minimization of the sum over examinees of squared differences between the target and transformed TCCs at given values of the latent variable (proficiency) in the IRT model. Using $\hat{P}_{jki}(\theta_i)$ to represent the target-group estimated probability (calculated using estimates of the item parameters) for the *kth* level of item *j* for a specific value of proficiency, θ_i, and $\hat{P}^*_{jki}(\theta_i)$ the probability estimate of the other group after the transformation (hence

[2]Note that Kolen and Brennan (2004) used parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for means and standard deviations, respectively), whereas I use statistics. Because actual linking procedures are carried out using sample data, I use statistical notation consistent with this practice. The transformation constants can, of course, be considered to be estimates of parameters defined using the population means and standard deviations.

incorporating the two transformation constants, K_1 and K_2), the Stocking-Lord method finds the transformation constants that minimize the expression

$$\sum_{i \in Q} w_i \left[\sum_{j=1}^{J} \sum_{k=1}^{m_j - 1} k\hat{P}_{jki}(\theta_i) - \sum_{j=1}^{J} \sum_{k=1}^{m_j - 1} k\hat{P}^*_{jki}(\theta_i) \right]^2, \tag{4.3}$$

where Q represents a set of values of the proficiency variable θ, J represents the number of anchor items, and the w_i are weights.

The set Q can be defined in a number of ways. It may include the values on the θ scale where the estimates have been found for the entire sample of examinees, or it may be a set of values on the scale at equal intervals between two extremes of the scale. The weights in the latter case would represent the densities of the distribution of proficiencies at the points on the scale. These densities can be determined from an assumed distribution (e.g., the normal) or from the distribution of sample estimates. Kolen and Brennan (2004) listed five ways of defining the points in Q. Note that the two terms within brackets in Equation 4.3 represent sums of values of the IRF for the jth item, before and after transformation. For example, the target IRF for item j is

$$IRF_{ji} = \sum_{k=1}^{m_j - 1} k\hat{P}_{jki}(\theta_i),$$

and the sum of the IRFs at θ_i is

$$\hat{\zeta}_i = \sum_{j=1}^{J} \sum_{k=1}^{m_j - 1} k\hat{P}_{jki}(\theta_i),$$

where $\hat{\zeta}_i$ represents the value of the sample TCC at θ_i which can be considered to be an estimate of the population value, ζ_i. Minimizing Equation 4.3 involves finding the K_1 and K_2 that minimize the weighted sum of squared differences between the target and transformed TCCs. Note that for a dichotomously scored item the IRF is simply the item characteristic curve.

4.3.1.3 The Haebara TCC Method

A method developed by Haebara (as cited in Kolen & Brennan, 2004) is an alternative to the Stocking-Lord method. This method defines the sum of squared differences between the IRFs of the common items across values on the scale and determines the values of K_1 and K_2 that minimize the sum of this quantity over examinees. The expression for the quantity minimized in this procedure is

$$\sum_{i \in Q} w_i \sum_{j=1}^{J} \left[\sum_{k=1}^{m_j - 1} \hat{P}_{jki}(\theta_i) - \sum_{k=1}^{m_j - 1} \hat{P}^*_{jki}(\theta_i) \right]^2.$$

Hence, the Haebara method finds the transformation constants that minimize the weighted sum (over items and proficiency values, θ_i) of squared differences between the two sets of IRFs, whereas the Stocking Lord method minimizes the weighted sum (over proficiency values) of squared differences between the two TCCs.

4.3.1.4 Concurrent Calibration

An alternative to the transformation methods described above is a concurrent calibration. This method, as described by Kolen (2006, pp. 176–177), entails using a multiple-group calibration program, enabling the development of a scale that allows for the expected differences in score distributions across grades on the vertical scale. All items at all grade levels are placed on the same vertical scale without the need for further transformations. The groups are the grade levels, and the common-item blocks across grades are critical to the scaling. Without those linking blocks, the results of the calibration analysis would be identical to separate calibration analyses at each grade level

4.3.2 The Equivalent-Groups Design

Developing a vertical scale using this design, as mentioned above, involves selecting randomly equivalent samples of examinees at each grade level. Separate calibration analyses are first conducted for each group of examinees at a specified grade level (hence separate analyses within each grade level for each of the equivalent groups). Referring to Table 4.2, the two randomly equivalent third-grade samples would be separately calibrated, yielding two independent sets of estimates of the item parameters in item Block 3B, and at the same time Blocks 3A and 4A would be calibrated on the same scale. The independent estimates for Block 3B could be averaged to provide the best estimates of those item parameters. The resulting means, however, may yield biased estimates. An alternative is concurrent calibration within grades. On the other hand, independent calibration followed by examination of differences in Block 3B IRFs of the two samples would be useful for studying model fit or sample equivalency issues. Similarly, the three independent equivalent samples at the fourth grade would be separately calibrated, yielding estimates of fourth-grade Blocks 4A and 4B as well as of Blocks 3B and 5A, all on the same scale. Then, the mean-mean, mean-sigma, Stocking-Lord, or Haebara method would be used to place the third- and fourth-grade items on the same scale. Similar analyses in a chain of linking analyses across the grade levels would result in the vertical scale across the seven grades. This part of the methodology is hence similar to that described above for the nonequivalent-groups designs. Alternatively, concurrent calibration procedures could be used to place all item parameters across all grade levels on the same vertical scale.

An alternative to the equivalent-groups design discussed above involves administering tests for two grade levels in randomly equivalent samples in the higher of the two grade levels. I will illustrate using Grades 3, 4, and 5 as an example. The first step is to administer Grade 3 item blocks to one fourth-grade sample and Grade 4 item blocks to a randomly equivalent fourth-grade sample. Then, use the mean-sigma procedure to align the estimates of proficiency (θ) to place the Grades 3 and 4 item parameter estimates on the same scale. In the second step, repeat this with Grades 4 and 5 item blocks administered to two randomly equivalent Grade 5 student samples. Finally, link the Grade 3–4 and Grade 4–5 scales using the Grade 4 items as an anchor set in a TCC method.

4.4 Model Fit Procedures

Model fit procedures can be used when conducting a vertical linking by both the common-item and equivalent-groups designs. A number of model fit procedures are available for assessing the calibration results of IRT scaling in general. Here I focus only on the methods used during vertical linking.

One of the most common procedures is to compare plots of the anchor item IRFs of each item for the two groups (equivalent-groups design, as mentioned above) or for the two forms (common-item design). In practice this procedure is usually limited to examination of the plots. Conceivably, however, IRT methods sometimes used to compare IRFs in the context of differential item functioning analyses could be used in the vertical scaling context. Although individual item parameter estimates could be compared, as mentioned above different sets of estimates can result in highly similar IRFs, and the similarity of the latter is most important in using this method to examine model fit.

Another methodology that is often used with designs involving anchor item sets is comparison of the three TCCs involved. To illustrate, consider that a scale has been established within the third grade in the common-item design of Table 4.1 and we are linking the fourth-grade scale to it. The three TCCs are of (a) the Grade 3 target data, (b) the untransformed Grade 4 data, and (c) the transformed Grade 4 data. In this example, the criterion of importance is that the transformed Grade 4 TCC be as identical as possible to the target Grade 3 TCC. Comparison of the untransformed with the transformed Grade 4 data simply provides information about how the transformation affected the TCC.

Another important aspect of the vertical scale development that should be examined is the progression of the scaling results across grade levels. One way of examining this is to plot the TCCs of each grade level as separate curves in a single plot. If the vertical scaling has been successful, the plot should show curves that do not cross, with an orderly progression of location of the curves across grades (lowest grade located lowest on the scale and each higher grade located somewhat higher than the next lower grade). The distances between these curves need not be

the same, because there is usually no basis for assuming that grade itself (i.e., Grade 3, Grade 4) is an interval scale.

In a typical testing program, different test forms are used within each grade in each assessment year. The new forms are usually equated within each grade level to the old forms. As the vertical scale is used across calendar years, the model-fit methodology results should be compared from year to year. Such comparison may reveal problems developing with the scale or the scaling procedures. Some items, for example, may show changes in the IRFs across years due to item parameter drift. If the exact same set of items is used year to year, drift can be detected through examination of the TCCs. If alternate forms are used each year, differences in the TCCs could reflect selection of differentially difficult items, but this normally would be taken care of through the within-grade year-to-year equating. Changes in the patterns of the TCCs across grades from year to year may be an indication of scale drift. If such things are observed, investigation through discussion with content experts and school officials should be undertaken to determine any curricular or population changes. In the event that these cross-year comparisons bring into question the validity of the scale, the scale may need to be reset by redoing the vertical linking with more recent data than used to develop the original scale.

4.5 Discussion

In this chapter I have described the most commonly used designs and methodology for developing vertical scales. Because the most common application is in large-scale educational assessment programs, the focus has been on methods used in such programs, primarily those using IRT models. There are many variations on the methodology discussed and alternative methodology not discussed in this chapter, so the reader is encouraged to refer to references cited herein as well as sources cited in those references. Additionally, I would like to point out that new research and development in this area is currently produced with some frequency, so those individuals wishing to keep current on the topics of this chapter should read the latest journals and conference programs in which psychometric methods are reported.

Author Note: Any opinions expressed in this chapter are those of the authors and not necessarily of Educational Testing Service.