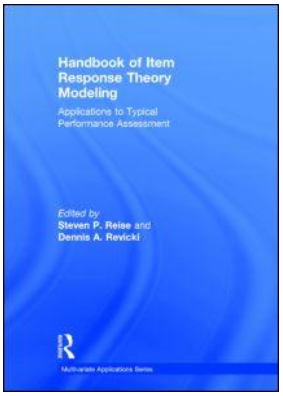


# Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

## Multidimensional Test Linking

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch19
Jonathan P. Weeks
**Published online on: 16 Dec 2014**

# 19 Multidimensional Test Linking

*Jonathan P. Weeks*

## Introduction

In order to compare scores from two or more related tests it is necessary for them to be reported on a common scale. Various linking methods have been developed for this purpose, depending on the statistical properties of the tests and the data collection design (cf., Dorans, Pommerich, & Holland, 2007; Kolen & Brennan, 2004); however, most of these methods are premised on the assumption that a single construct is measured within and between tests. When a multidimensional model is applied to tests measuring more than one dimension it is important to consider the comparability of dimension-specific scores. The goal of this chapter is to provide a foundation for understanding multidimensional test score linking and various issues that should be considered in this process.

This chapter is organized into four main parts. The first part provides an overview of score linking within a unidimensional item response theory (IRT; Lord & Novick, 1968) framework and the correspondence to Procrustean transformations (Gower & Dijksterhuis, 2004; Mulaik, 1972) that serve as the basis for multidimensional test score linking. The second part presents the equations for score linking within a multidimensional item response theory (MIRT; Reckase, 2009) framework. This is followed by a section that lays out considerations for when it may or may not be appropriate to link scores multidimensionally. The final part provides a discussion of properties that should be evaluated as part of the linking process; these properties are addressed using empirical data from a large-scale mathematics assessment.

## Research Methods

Methods for linking score scales have been around for more than a century (cf., Hull, 1922; Kelley, 1923). These methods include equipercentile equating (Braun & Holland, 1982; Kelley, 1923), linear raw-score methods (Angoff, 1971; Gulliksen, 1950; Kolen & Brennan, 2004; Levine, 1955), and IRT-based methods (Kim & Lee, 2006; Kolen & Brennan, 2004), among others. The focus of this chapter is primarily on linear transformations of score scales within an IRT framework when the item parameters for the tests of interest are estimated separately. Within this framework, the goal of the linking is to adjust the scores from different test forms so that the scores can be compared on a common scale and the forms can be used interchangeably. That is, the scores from all tests of interest must be transformed to the metric of a reference test (or a reference scale). The methods used to transform the scores are based principally on two characteristics: the equivalence (or lack thereof) of the examinee groups and the associated score distributions, and the statistical and construct-related properties of the test forms. The equivalence of the examinee groups is typically related to the data collection design while the properties of the tests are

related to the various aspects of the test design (e.g., the construct of interest, difficulty, item format, etc.).

## Data Collection Designs and Terminology

Three of the most common data collection designs (Kolen & Brennan, 2004) are the 1) randomly equivalent groups, 2) single-group with counterbalancing, and 3) nonequivalent groups common item design. In a randomly equivalent groups design each examinee group takes a different test form. The design assumes that the underlying distributions of true scores are equivalent for each group; hence, differences in observed performance can be attributed to differences in form difficulty. By adjusting the scores on each form to account for these differences, the scores across forms can be compared on a common scale. In a single-group design, the same examinees take each form. To minimize order effects in the linking, the administration of forms is typically counterbalanced (e.g., half of the examinees take Form X first and Form Y second while the other half takes Form Y first and Form X second). The transformation of scores then proceeds using the same methods employed for a randomly equivalent groups design.[1] When the groups are not equivalent—when the examinees come from different populations—differences between the tests and groups must be simultaneously considered (i.e., disentangling how much of the difference in scores is due to test difficulty versus group ability). In these instances, a nonequivalent groups common item design may be used to anchor the tests together. Under this design a given subset of items is administered on both tests and the relationship between performance on these common items across groups is used as the basis for linking the scores.

A variety of methods have been developed for linking test scores that are related in part to the associated data collection design; however, the terms used to differentiate between these methods are typically tied to the characteristics of the tests. Throughout this chapter I use three terms seemingly interchangeably (*linking, scaling*, and *equating*); however, these terms denote various assumptions with respect to the comparability of the constructs measured by each test as well as associated statistical specifications. My use of terms is generally consistent with those presented by Kolen (2004). *Linking* is a generic term for any approach used to make results comparable. *Scaling*[2] is a more restrictive term used to denote the linking of scores for tests that measure the same construct, yet differ with respect to test specifications (e.g., differences in content coverage or reliability). *Equating* is a restricted form of scaling where the test forms are constructed using the same test specifications and satisfy a set of specific criteria (addressed in the last part of this chapter).

## Unidimensional Test Score Linking

Most of the methods developed for test score linking are premised on the assumption of unidimensionality (cf., Dorans, Pommerich, & Holland, 2007; Kolen & Brennan, 2004; Skaggs & Lissitz, 1986); that is, the assumption that the scores across test forms characterize examinee performance on a single construct. The primary focus of this chapter is

1 Because of practical considerations like test time and cost, the single-group design is rarely used.

2 In Kolen (2004), the term *calibration* is used instead of *scaling;* however, the term *calibration* is often used within an IRT context to distinguish between the simultaneous estimation of item parameters for multiple groups (concurrent calibration) and the estimation of item parameters for each test separately (separate calibration). To reduce confusion, the term *scaling* is used as the restricted form of linking.

on multidimensional score linking; however, the methods developed for unidimensional linking provide an important context for understanding linking within a multidimensional framework. The following section provides an overview of unidimensional IRT-based test score linking within a separate calibration framework. Readers familiar with these methods may wish to skip this section and jump to the discussion of Procrustean transformations that serve as the basis for multidimensional test score linking.

In item response theory, an examinee's response to a test item is modeled probabilistically as a function of the test taker's latent ability and the item's characteristics. Let the variable $X_{ij}$ represent the response of examinee $j$ to item $i$. Given a test consisting of dichotomously scored items, let $X_{ij} = 1$ for a correct item response and $X_{ij} = 0$ for an incorrect response. The item response curve for the two-parameter logistic model (2PL; Birnbaum, 1968) takes the following form:

$$P\left(X_{ij} = 1 \mid \theta_j, a_i, b_i\right) = \frac{exp\left(Da_i\left[\theta_j - b_i\right]\right)}{1 + exp\left(Da_i\left[\theta_j - b_i\right]\right)}, \tag{19.1}$$

where $\theta_j$ is an individual's latent ability on a single construct, $a_i$ is the item discrimination (slope), $b_i$ is the item difficulty, and $D$ is a scaling constant. When items are polytomously scored (i.e., items with three or more score categories, as with constructed response items), the response $X_{ij}$ can be coded using a set of responses $k = \{0, \ldots, K_i - 1\}$ where $K_i$ is the total number of categories for the item. When the values of $k$ correspond to successively ordered categories, the response probabilities can be modeled using a model like the generalized partial credit model (GPCM; Muraki, 1992), which takes the following form:

$$P\left(X_{ij} = k \mid \theta_j, a_i, b_{ik}\right) = \frac{exp\left[\sum_{v=1}^{k} Da_i\left(\theta_j - b_{iv}\right)\right]}{\sum_{h=1}^{K_i} exp\left[\sum_{v=1}^{h} Da_i\left(\theta_j - b_{iv}\right)\right]}. \tag{19.2}$$

where $b_{ik}$ is a step-intersection parameter. The other parameters have the same interpretation as in the 2PL. Other models such as the three-parameter logistic model (3PL; Birnbaum, 1968), Rasch model (1960), and various polytomous models can be specified, but in many cases the linking equations are the same as those presented below for the 2PL/GPCM. For a more complete explication of unidimensional IRT-based linking methods for dichotomous and polytomous models, see Kolen and Brennan (2004) or Kim and Lee (2006).

The use of IRT in general, and for score linking, is premised on two strong, related, assumptions: local independence and unidimensionality. A set of items is considered locally independent if, for fixed values of the underlying construct, the item responses are statistically independent. In order for this to hold, the dimensional structure must be adequately specified within and across tests. More specifically, all of the items on the tests must measure a single construct. In the context of score linking, there is an added assumption that the item parameters are unbiased (e.g., that there is no differential item functioning between subpopulations). When these assumptions are met, and when the specified model fits the data, many IRT models share a property that makes IRT well suited for test linking: parameter invariance. That is, the parameters for a given item should be the same regardless of the subpopulation used to estimate them, and the ability for a given examinee should be the same regardless of the test administered.

Scaling and equating within an IRT framework is typically based on a common item design[3] that assumes invariance of the common item parameters across examinee groups and tests. When the item parameters for two or more tests are estimated concurrently using a multigroup approach (Bock & Zimowski, 1997; von Davier & von Davier, 2007), the common item parameters are usually constrained to be equal across groups while the latent ability distributions for each of the groups are allowed to vary (the mean and standard deviation of a reference group are typically fixed for identification purposes). This approach places all of the item parameter estimates and subsequent estimates of examinee abilities on a common scale. As an alternative to this approach, item parameters can be estimated separately for each test and then linked via a secondary process.

In the separate calibration approach the mean and standard deviation of the latent ability distributions are often constrained (e.g., to zero and one respectively) when estimating the item parameters. The linking then proceeds by using a linear scaling function to identify a set of coefficients that minimize the differences between common item parameter estimates. These linking coefficients are used to transform all of the item parameter and ability estimates for the focal tests and place them on the scale of the reference test. In cases where it is necessary to link scores from three or more forms, a chain-linking process (i.e., a series of linear transformations) can be used (Kolen & Brennan, 2004); however, additional error may be introduced via successive transformations. As an alternative to chain linking, a set of linking coefficients can be simultaneously estimated to transform the scores from each form directly to the reference scale (Haberman, 2009). Von Davier and von Davier (2007) show that when the various IRT assumptions are met, concurrent calibration is preferable to separate calibration insomuch that it reduces error in the linking; however, to provide a more explicit connection to various multidimensional linking methods, the separate calibration approach is described in more detail below.

With a separate calibration approach, the means and standard deviations of the scores on one or more tests are adjusted to resolve any differences between the test forms. A simple linear equation,

$$\theta_T = A\theta_F + B,$$

can be used to transform $\theta_F$ values to the $\theta_T$ scale (the subscripts F and T correspond to what I will refer to as the *from* scale and *to* scale respectively) where $A$ and $B$ are estimated coefficients that adjust the scale (variability) and location (mean) of the scores respectively. Because the item parameters are defined with respect to a $\theta$ scale, any transformation of the scale will necessarily require a change in the item parameters so that the expected response probabilities remain unchanged. It can be readily shown that $a_i$ and $b_{ik}$ for the 2PL and GPCM on the *from* scale can be transformed to the *to* scale by (Baker, 1992; Kim & Lee, 2006; Lord & Novick, 1968):

$$\begin{aligned} a_{iT} &= a_{iF} / A \\ b_{ikT} &= Ab_{ikF} + B . \end{aligned} \tag{19.3}$$

In practice, the parameters $\theta$, $a_i$ and $b_{ik}$ for each test are unknown; hence, estimates of the parameters ($\hat{\theta}$, $\hat{a}_i$ and $\hat{b}_{ik}$ respectively) must be considered with respect to the linking.

3 Note that if a randomly equivalent group design is employed, the estimation can be run as a single group model or, in the case of separate calibration, the same identification constraints can be used for the latent ability distribution so that no subsequent linking is required (Kolen & Brennan, 2004).

Based on the assumption of parameter invariance, the goal is to find a set of coefficients that minimize the difference between the untransformed *to* scale and transformed *from* scale common item parameter estimates. Several approaches have been developed for this purpose. The mean/sigma (Marco, 1977) and mean/mean (Loyd & Hoover, 1980) methods, commonly known as moment methods, are the simplest approaches to estimating *A* and *B* because they only require the computation of means and standard deviations for various common item parameter estimates. For mean/sigma, only the difficulty parameters are used. That is,

$$A = \frac{\sigma\left(\hat{b}_T\right)}{\sigma\left(\hat{b}_F\right)}$$
$$B = \mu\left(\hat{b}_T\right) - A\left[\mu\left(\hat{b}_F\right)\right], \tag{19.4}$$

where $\mu(\,)$ and $\sigma(\,)$ are the means and standard deviations, taken over all $C \leq J$ common items and $K_c$ response categories. One potential limitation of this approach is that it does not consider the slope parameters. The mean/mean, on the other hand, uses both the slope and difficulty parameters to estimate the linking coefficients where:

$$A = \frac{\sigma\left(\hat{a}_F\right)}{\sigma\left(\hat{a}_T\right)}$$
$$B = \mu\left(\hat{b}_T\right) - A\left[\mu\left(\hat{b}_F\right)\right]. \tag{19.5}$$

As an alternative to the moment methods, Haebara (1980) and Stocking and Lord (1983) developed characteristic curve methods that use an iterative approach to estimate the linking coefficients by minimizing the sum of squared differences between item characteristic curves (ICC) and test characteristic curves (TCC) for the common items for the two methods respectively. The Haebara method minimizes:

$$Q = \sum_{g=1}^{G}\sum_{c=1}^{C}\sum_{k=1}^{K_c}\left[P_{ck}\left(\theta_g\right) - P^*_{ck}\left(\theta_g\right)\right]^2, \tag{19.6}$$

while the Stocking-Lord method minimizes:

$$F = \sum_{g=1}^{G}\left[\sum_{c=1}^{C}\sum_{k=1}^{K_c}U_{ck}P_{ck}\left(\theta_g\right) - \sum_{c=1}^{C}\sum_{k=1}^{K_c}U_{ck}P^*_{ck}\left(\theta_g\right)\right]^2. \tag{19.7}$$

The $\theta_g$ are a set of G points on the *to* scale where differences in expected probabilities are evaluated, $P_{ck}\left(\theta_g\right)$ are expected probabilities based on the untransformed *to* scale common item parameter estimates, and $P^*_{ck}\left(\theta_g\right)$ are expected probabilities based on the transformed *from* scale common item parameter estimates.[4] To create the test characteristic curves

4 Extensions of the Haebara and Stocking-Lord method can be specified that consider a symmetric minimization of the criterion with or without quadrature weights (see Kim & Lee, 2006 for more information).

for the Stocking-Lord method, the scoring function $U_{ck}$ must be included (particularly for polytomous items). These values are usually specified as $U_{ck} = \{0,\ldots,K_c - 1\}$ which assumes that the categories are ordered.

## Procrustean Transformations

Multidimensional test score linking has its origins in Procrustean transformations (cf., Gower & Dijksterhuis, 2004; Mulaik, 1972). Simply put, the goal of Procrustean methods is to transform the values in a given matrix (e.g., a matrix of factor loadings or item slopes), $\boldsymbol{X}_F(N \times P_1)$, so that they align as closely as possible to values in a target matrix, $\boldsymbol{X}_T(N \times P_2)$. The simplest formulation of a Procrustean transformation is the minimization of:

$$H = \| \boldsymbol{X}_F\boldsymbol{T} - \boldsymbol{X}_T \|, \tag{19.8}$$

where $\boldsymbol{T}$ is a $P_1 \times P_2$ transformation matrix and the operator $\{\| \|\}$ is the matrix norm defined by the sums of squares.[5] Different variants of the transformation can be specified depending on the dimensions of $\boldsymbol{X}_F$ and $\boldsymbol{X}_T$ and any restrictions on the form of $\boldsymbol{T}$. The particular variations considered in this chapter are cases where $\boldsymbol{T}$ is an oblique transformation or an orthogonal rotation. In the latter case, the transformation will be denoted by $\boldsymbol{Q}$. The transformation matrices $\boldsymbol{T}$ and $\boldsymbol{Q}$ play one or two roles in the minimization of the difference between $\boldsymbol{X}_F$ and $\boldsymbol{X}_T$. Both matrices adjust for rotational indeterminacy in order to orient the values along the same axes within the multidimensional space, yet with the oblique transformation, $\boldsymbol{T}$, there is also an intrinsic scaling that adjusts the variability of each dimension in $\boldsymbol{X}_F$. With the orthogonal rotation, $\boldsymbol{Q}$, there are no adjustments made to the variability; hence, a separate scaling (dilation) matrix may be needed. In the latter case, the criterion to be minimized is:[6]

$$H = \| \boldsymbol{X}_F\boldsymbol{R}\boldsymbol{Q} - \boldsymbol{X}_T \| \text{ or } H = \| \boldsymbol{X}_F\boldsymbol{Q}\boldsymbol{S} - \boldsymbol{X}_T \|. \tag{19.9}$$

The matrices $\boldsymbol{R}$ and $\boldsymbol{S}$ correspond to pre-scaling and post-scaling transformations respectively. These are diagonal matrices of scaling coefficients. If the variability of each dimension is adjusted by a single constant, this is referred to as isotropic scaling; if the variability is adjusted differently for each dimension, this is referred to as anisotropic scaling. The final type of transformation that may be required is a translation of the values in $\boldsymbol{X}_F$. In this case, when no additional scaling matrix is specified, the criterion to be minimized is:

$$H = \|\boldsymbol{X}_F\boldsymbol{T} - \boldsymbol{1}\boldsymbol{a}' - \boldsymbol{X}_T \|, \tag{19.10}$$

where $\boldsymbol{a}' = \boldsymbol{a}'_F T - \boldsymbol{a}'_T$. The translation vector $a$ essentially shifts the location of the means of each scale. There is no *a priori* defined sequence to multivariate transformations; that is,

5 Estimation of the various elements for rotation, scaling, and translation are presented in the following section in the discussion of test linking within a multidimensional item response theory framework.

6 It should be noted that a constrained oblique transformation for $\boldsymbol{T}$ can be specified with a pre-scaling and/or post-scaling transformation; however, for the purpose of this chapter only orthogonal transformations with an associated scaling matrix are considered. See Gower and Dijksterhuis (2004) for information on additional scaling transformations.

the sequence of rotation, scaling, and translation may occur in any order by restructuring the formulation of the minimization function accordingly. As such, a decision should be made regarding the sequence of transformations and applied consistently in any practical application of multidimensional linking.

The transformations presented above provide a simplified overview of the underpinnings of multidimensional test score linking, yet there are two other important considerations that can be addressed using Procrustean methods: the choice of a general reference scale and projection problems. The simplest formulation of the Procrustes problem seeks a solution to transform $\boldsymbol{X}_F$ onto the scale of $\boldsymbol{X}_T$ (a one-sided Procrustes problem); however, there may be instances where it is desirable to transform $\boldsymbol{X}_F$ and $\boldsymbol{X}_T$ to an "average" configuration. This is referred to as a two-sided Procrustes problem. The criterion to be minimized for two groups takes the following form:

$$H = \|\boldsymbol{X}_1\boldsymbol{T}_1 - \boldsymbol{X}_2\boldsymbol{T}_2\|. \tag{19.11}$$

Because there is no longer a *to* scale and *from* scale, the subscripts are denoted by numbers. Notice also that instead of a single transformation matrix, each $\boldsymbol{X}$ has an associated $\boldsymbol{T}$. To illustrate the transformation to an "average" reference matrix, $\boldsymbol{G}$, the two-sided criterion can be reformulated as a generalized Procrustes problem (Gower & Dijksterhuis, 2004):

$$H = K\sum_{k=1}^{K}\| \boldsymbol{X}_k\boldsymbol{T}_k - \boldsymbol{G} \|, \tag{19.12}$$

where:

$$\mathbf{G} = K^{-1}\sum_{k=1}^{K}(\mathbf{X}_k\mathbf{T}_k). \tag{19.13}$$

The two-sided approach is particularly relevant for multidimensional score linking in the implementation of oblique transformations. In order to obtain an oblique transformation that is symmetric, the two-sided criterion is minimized to estimate a single $\boldsymbol{T}$ (for a pair of tests) rather than $\boldsymbol{T}_k$.[7] On the other hand, it can be shown that the one-sided and two-sided approach in the estimation of $\boldsymbol{Q}$ both provide an orthogonal rotation. Hence, the simpler one-sided approach can be used to estimate $\boldsymbol{Q}$. As a further extension of the two-sided Procrustes problem, a generalized Procrustes approach could be used to link more than two test forms, although this approach has not been fully explored in the literature on multidimensional test score linking.

Another consideration in multidimensional test score linking is the number of dimensions underlying each test. The equations given earlier suggest that the dimensions of $\boldsymbol{X}_F$ and $\boldsymbol{X}_T$ are $(N\times P_1)$ and $(N\times P_2)$ respectively where $P_1 = P_2$. However, there may be instances where there are, say, four factors associated with $\boldsymbol{X}_F$ and three factors associated with $\boldsymbol{X}_T$ or, say, three factors associated with both $\boldsymbol{X}_F$ and $\boldsymbol{X}_T$ but with only two common factors. These scenarios may occur in K-12 assessments where tests are administered at different grade levels and it is possible that the measured constructs shift somewhat from grade to grade (this issue is discussed in more detail later). Some researchers may argue that

7 This is consistent with the notion of symmetric transformations in unidimensional linking presented by Haebara (1980).

linking scores on tests like these is not defensible while others may support the approach for the purpose of creating a single score scale for the common dimensions across tests.

When $P_1 \neq P_2$, $\boldsymbol{T}$ will not be a square transformation matrix. This is a projection Procrustes problem (Gower & Dijksterhuis, 2004). Different methods have been developed to handle cases where higher-dimensional matrices are projected onto lower-dimensional spaces and vice versa, but for the present discussion it can be shown that the correct solution for $\boldsymbol{T}$ and $\boldsymbol{Q}$ can be obtained by padding "missing" columns in $\boldsymbol{X}_F$ and/or $\boldsymbol{X}_T$ with zeros. For instance, with respect to the first example, $\boldsymbol{X}_F$ would be an $(N \times 4)$ matrix where all four columns include item slopes and $\boldsymbol{X}_T$ would be an $(N \times 4)$ matrix where the first three columns include item slopes and the fourth column includes all zeros.

There are number of additional constraints and/or extensions related to the transformations presented above that one might consider in order to address specific issues in multidimensional test linking, but it is beyond the scope of this chapter to delve into these constraints/extensions in more depth. The transformations discussed here represent key components that connect the linear transformations used in unidimensional test score linking to a multidimensional conceptualization of score linking. The reader interested in a more complete explication of Procrustes problems should consult Gower and Dijksterhuis (2004).

## Multidimensional Test Score Linking

In the previous section, Procrustean methods were presented as the foundation upon which multidimensional test score linking is built (in a separate calibration context). In this section, extensions of the Procrustean transformations are provided for linking scores within a multidimensional item response theory framework. As in the unidimensional case, let the variable $X_{ij}$ represent the response of examinee $j$ to item $i$. Given a test consisting of dichotomously-scored items, let $X_{ij} = 1$ for a correct item response, and let $X_{ij} = 0$ for an incorrect response. The item response function for the multidimensional two-parameter logistic model (M2PL; Reckase, 1985) takes the following form:

$$P\left(X_{ij} = 1 \mid \boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i\right) = \frac{exp\left(\boldsymbol{a}_i' \boldsymbol{\theta}_j + d_j\right)}{1 + exp\left(\boldsymbol{a}_i' \boldsymbol{\theta}_j + d_j\right)}, \tag{19.14}$$

where $\boldsymbol{\theta}_j$ is a vector of latent abilities in M dimensions, $\boldsymbol{a}_i$ is a vector of item slopes, and $d_i$ is a scalar parameter related to item difficulty. When the item responses correspond to successively ordered categories, the response probabilities can be modeled using a model like the multidimensional generalized partial credit model (MGPCM; Yao & Schwarz, 2006), which takes the following form:

$$P\left(X_{ij} = k \mid \boldsymbol{\theta}_j, \boldsymbol{a}_i, d_{ik}\right) = \frac{exp\left[\sum_{v=1}^{k} \boldsymbol{a}_i' \boldsymbol{\theta}_j + d_{iv}\right]}{\sum_{h=1}^{K_i} exp\left[\sum_{v=1}^{h} \boldsymbol{a}_i' \boldsymbol{\theta}_j + d_{iv}\right]}, \tag{19.15}$$

where $d_{ik}$ is a step-intersection parameter. The other parameters have the same interpretation as in the M2PL. As in the unidimensional case, other multidimensional models can be specified for dichotomous and polytomously scored responses; however, the M2PL/MGPCM will be used here because it aligns closely with the frequently used *common factor model* in factor analysis (Thurstone, 1931, 1935, 1947).

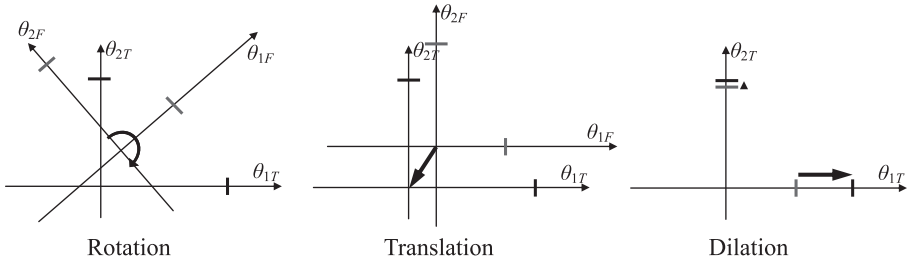


*Figure 19.1* Transformations in multidimensional linking.

In order to establish a multidimensional scale across two or more forms, the item parameters can be estimated concurrently or separately. In the concurrent case, an extension of the multigroup method presented by Bock and Zimowski (1997) and von Davier and von Davier (2007) could be used where the common item parameters are constrained to be equal across groups. The latent ability distribution for a given group is usually specified as a reference group where the means and variances are fixed (typically with means equal to zero and variances equal to unity). Depending on the specified factor structure (e.g., an exploratory versus confirmatory structure), additional constraints may also be placed on the covariances for the reference group. The means, variances, and covariances for the remaining groups can be freely estimated. The end result of the concurrent calibration is that all of the item parameter and ability estimates will be on a common scale. As an alternative to the concurrent approach, the item parameters and examinee abilities can be estimated separately for each group. The linking then proceeds by finding linking coefficients that can be used to adjust for differences in the common item parameter estimates (or estimated latent ability distributions) for each group. In short, the goal is to implement a Procrustean transformation to place the results on a common scale. Building on the transformations described in the previous section, an explication of multidimensional separate calibration methods is presented below.

When conducting a separate calibration in the unidimensional case, a set of linking coefficients is estimated to adjust the location and scale of the ability distribution for the test being transformed. A corresponding transformation of the item parameters is made to maintain the expected probabilities of a given response. Similarly, in the multidimensional case, a set of coefficients can be estimated to adjust the scale and location for each of the modeled dimensions; however, a rotation may also be needed to align the dimensional axes. Figure 19.1 is a graphical representation of the various elements that must be resolved for linking tests in two dimensions;[8] the same premise holds for linking tests in more than two dimensions. In this figure, the axes with the labels $\theta_{1T}$ and $\theta_{2T}$ correspond to the *to* scales. The other set of axes correspond to the *from* scales. The goal of the linking process, conceptually, is to adjust the axes for the test being transformed so that they lie perfectly on top of the set of reference axes (i.e., adjusting for differences between the common item parameters). In general, this is a three-step process. The first step involves rotating the axes so that they are parallel. Once the axes are aligned, the *from* axes must be shifted so that the crosshairs lie on top of the *to* axes (translation). This is equivalent to adjusting the mean of the *from* distribution in a unidimensional context. The final step involves stretching or shrinking the scale associated with each axis so that

8 This illustration presumes an orthogonal transformation; however, the general premise of rotation, translation, and dilation holds for oblique transformations as well.

the perpendicular bars (representing the variability of each scale) lie on top of one another (scaling or dilation). This is equivalent to adjusting the standard deviation of the *from* scale in the unidimensional context.

## MIRT Linking Equations

Several approaches have been developed to link tests on multiple dimensions with the key distinction being the rotation method (orthogonal vs. oblique) and dilation method (isotropic vs. anisotropic). In general, the parameters on the *from* scale can be transformed to the *to* scale via (Oshima, Davey, & Lee, 2000; Yao & Boughton, 2009):

$$\begin{aligned} \boldsymbol{a}'_{iT} &= \boldsymbol{a}'_{iF}\boldsymbol{T}^{-1} \\ d_{ikT} &= d_{ikF} - \boldsymbol{a}'_{iF}\boldsymbol{T}^{-1}\mathbf{m} \\ \boldsymbol{\theta}_{jT} &= \mathbf{T}\boldsymbol{\theta}_{jT} + \mathbf{m}, \end{aligned} \tag{19.16}$$

where $\boldsymbol{T}$ is an oblique transformation matrix that adjusts for rotational indeterminacy and differences in variability, and $\mathbf{m}$ is a translation vector that adjusts for differences in difficulty on each dimension. Alternatively, the parameters can be transformed using an orthogonal rotation, $\boldsymbol{Q}$, and an anisotropic scaling matrix, $\boldsymbol{S}$ (Min, 2007):

$$\begin{aligned} \boldsymbol{a}'_{iT} &= \boldsymbol{a}'_{iF}\boldsymbol{Q}^{-1}\boldsymbol{S}^{-1} \\ d_{ikT} &= d_{ikF} - \boldsymbol{a}'_{iF}\boldsymbol{Q}^{-1}\mathbf{m} \\ \boldsymbol{\theta}_{jT} &= \mathbf{S}\left(\mathbf{Q}\boldsymbol{\theta}_{jT} + \mathbf{m}\right). \end{aligned} \tag{19.17}$$

As an added constraint to this approach, $\boldsymbol{S}$ can be specified as a diagonal isotropic scaling matrix with a single dilation parameter (Li & Lissitz, 2000). This supposes that the variability of each dimension is adjusted by a constant value. It is useful to note that when the data are modeled using a between-item dimensional structure where each item only loads on a single dimension, $\boldsymbol{Q}$ will be an identity matrix and $\boldsymbol{S}$ will equal $\boldsymbol{T}$.

As in the unidimensional case, the item parameters and linking coefficients are unknown and must be estimated. The simplest approach to estimating $\boldsymbol{T}$ is via unconstrained least squares:

$$\mathbf{T} = \left(\hat{\boldsymbol{a}}'_F\hat{\boldsymbol{a}}_F\right)^{-1}\hat{\boldsymbol{a}}_F\hat{\boldsymbol{a}}_T, \tag{19.18}$$

where $\hat{\mathbf{a}}$ is the matrix of common item parameter estimates for $\mathbf{a}$. In subsequent equations, $\hat{d}_{ik}$ is an estimate of $d_{ik}$. In general, the linking coefficients obtained via unconstrained least squares regression do not typically satisfy the symmetry property of equating (Lord, 1980); hence, additional constraints must be applied to the estimation of $\boldsymbol{T}$ to make it a symmetric transformation. In order to maintain the symmetry property, as well as the oblique transformation, $\boldsymbol{T}$ can be estimated using a two-sided Procrustes approach via alternating least squares (Gower & Dijksterhuis, 2004). On the other hand, $\boldsymbol{T}$ can be constrained to be orthogonal. In this case, a singular value decomposition of $\hat{\boldsymbol{a}}'_F\hat{\boldsymbol{a}}_T$:[9]

$$SVD\left(\hat{\boldsymbol{a}}'_F\hat{\boldsymbol{a}}_T\right) = \boldsymbol{U}\,\boldsymbol{\Sigma}\boldsymbol{V}', \tag{19.19}$$

9 It is recommended that the values for $\boldsymbol{X}_F$ and $\boldsymbol{X}_T$ be mean centered prior to estimating the linking coefficients (Gower & Dijksterhuis, 2004; Schönemann & Carroll, 1970).

can be used to obtain the rotation matrix:

$$\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{U}' . \tag{19.20}$$

With orthogonal rotations it is possible for **Q** to produce a reflection rather than a pure rotation. This can be determined by examining [det($\boldsymbol{U}$) det($\boldsymbol{V}$)] where det is the determinant. If this value is negative, the sign in the final column of $\boldsymbol{U}$ or $\boldsymbol{V}$ can be changed to produce a rotation (Gower & Dijksterhuis, 2004). The diagonal matrix of anisotropic scaling coefficients can then be estimated as:

$$\mathbf{S} = \frac{diag\left(\boldsymbol{Q}\,\hat{\boldsymbol{a}}'_F\hat{\boldsymbol{a}}_T\right)}{diag\left(\boldsymbol{Q}\,\hat{\boldsymbol{a}}'_F\hat{\boldsymbol{a}}_F\boldsymbol{Q}^{-1}\right)} . \tag{19.21}$$

If the dilation matrix is isotropic, the scaling coefficient, *s*, can be estimated as:

$$\mathrm{s} = \frac{trace\left(\boldsymbol{Q}\boldsymbol{a}'_F\,\boldsymbol{a}_T\right)}{trace\left(\boldsymbol{a}'_F\,\boldsymbol{a}_F\right)} . \tag{19.22}$$

The translation vector **m** for both the oblique and orthogonal transformations can be estimated as:

$$\mathbf{m} = \left(\hat{\boldsymbol{a}}'_F\hat{\boldsymbol{a}}_F\right)^{-1}\hat{\boldsymbol{a}}_F\left(\hat{\boldsymbol{d}}_{ikF} - \hat{\boldsymbol{d}}_{ikT}\right) . \tag{19.23}$$

In all of the equations used to estimate the multidimensional linking coefficients, only the item parameter estimates for the $C \leq J$ common items are used. These coefficients are then used to transform all of the item and ability parameter estimates to the reference scale.

In addition to the least squares methods presented earlier, researchers have developed multidimensional extensions of the unidimensional moment and characteristic curve methods (cf., Li & Lissitz, 2000; Min, 2007; Oshima, Davey, & Lee, 2000; Reckase & Martineau, 2004). Oshima and colleagues (2000) and Yao and Boughton (2009) compared multidimensional extensions of the moment and characteristic curve methods and found that the characteristic curve methods perform better than the moment methods in all cases. Li and Lissitz (2000) compared the performance of the multidimensional Stocking-Lord and a constrained least squares approach under four equating and vertical scaling conditions. They found that in all cases the least squares approach resulted in the smallest amount of linking error, particularly in the context of vertical scaling. Based on these findings, estimation of the multidimensional linking coefficients via constrained least squares is recommended.

## When Is Multidimensional Score Linking Appropriate?

In practical applications of test linking, it is commonly assumed that all of the tests being linked measure a single construct; yet in many instances there are strong theoretical and empirical reasons to suspect that the construct of interest is multidimensional. Take, for example, subject areas like reading, mathematics, and science. A number of studies have

been conducted to examine cognitive and content dimensions within these domains (cf., Abedi, 1994; Davis, 1972; Embretson & Wetzel, 1987; Geary, 2006; Gierl, Tan, & Wang, 2005; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Kupermintz & Snow, 1997; Reckase & Martineau, 2004). In many cases, evidence of multidimensionality is present for both major and minor dimensions; however, almost all operational assessments of reading, mathematics, and science (e.g., K–12 state tests within the United States) treat the measured construct as unidimensional. There are a variety of explanations for this, but a key takeaway message is that evidence of underlying multidimensionality alone may be insufficient to justify the development of a multidimensional scale (or the subsequent linking of test scores on two or more dimensions). Other criteria such as reliability and interpretability must also be considered. The following section addresses subscore reliability, as it relates to the relative value of subscores versus total scores; score interpretability, as it pertains to construct shift; and common item requirements, as it relates to linking stability, in order to provide a context for when multidimensional test linking is appropriate.

## Subscore Reliability

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) state that "Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established." This standard applies to subscores as well as total scores. In the simplest case, raw subscores may be computed based on subsets of items within a given test. Alternatively, augmented subscores (Wainer et al., 2001) could be computed or factor scores/ability estimates from a multidimensional latent variable model could be estimated. An important consideration in all of these cases is whether the derived subscores are sufficiently reliable to be reported for individuals. In order to address this question, Haberman (2008) proposed a statistical method based on classical test theory (CTT) to evaluate whether subscores have added value relative to total scores.[10] The method is premised on the notion that subscores with added value are highly reliable and have correlations with the other subscores that are not very high. The approach relies on different regression-based estimates of true subscores and a comparison of the associated mean square error terms.

Consider an examinee, $j$, with a total raw score, $S_j$, and a raw subscore, $S_{jk}$, for skill $k$. The true score, $T_j$, associated with $S_j$ is the average score for the examinee over repeated administrations of the same test or parallel forms of the test. Similarly, $T_{jk}$ is the true subscore associated with $S_{jk}$. Haberman (2008) used three approaches to obtain estimates of the true subscores and the true subscore variance.

- $U_{jks} = \alpha_{ks} + \beta_{ks} S_{jk}$ is an estimate based on the raw subscore $S_{jk}$. This yields the following mean squared error: $\tau_{ks}^2 = E\left(\left[T_{jk} - U_{jks}\right]^2\right)$.
- $U_{jkx} = \alpha_{kx} + \beta_{kx} S_j$ is an estimate based on the raw total score $S_j$. This yields the following mean squared error: $\tau_{kx}^2 = E\left(\left[T_{jk} - U_{jkx}\right]^2\right)$.
- $U_{jkc} = \alpha_{kc} + \beta_{k1c} S_j + \beta_{k2c} S_{jk}$ is an estimate based on the raw total score $S_j$ and the raw subscore $S_{jk}$. This yields the following mean squared error: $\tau_{kc}^2 = E\left(\left[T_{jk} - U_{jkc}\right]^2\right)$.

10 This method also holds in the consideration of augmented subscores (Wainer et al., 2001).

To compare the possible subscores, the proportional reduction in mean squared error (PRMSE) is considered relative to the variance of the true raw subscore $\tau_{k0}^2 = E\left(\left[T_{jk} - E\left(T_{jk}\right)\right]^2\right)$. The PRMSEs for each subscore are:

$$\begin{aligned} PRMSE_{ks} &= 1 - \tau_{ks}^2 / \tau_{k0}^2 \\ PRMSE_{kx} &= 1 - \tau_{kx}^2 / \tau_{k0}^2 \\ PRMSE_{kc} &= 1 - \tau_{kc}^2 / \tau_{k0}^2 . \end{aligned} \tag{19.24}$$

It is important to note that $PRMSE_{ks}$ is the reliability of $S_{jk}$. All of the PRMSE values range from zero to one, with values near one being more desirable. When $PRMSE_{ks}$ is less than $PRMSE_{kx}$, the subscore provides little added value relative to the total score. On the other hand, if $PRMSE_{ks}$ is greater than $PRMSE_{kx}$, the subscores provide more diagnostic information than the total score. Haberman suggests that $PRMSE_{kc}$ should only be used when it is substantively larger than the maximum of $PRMSE_{ks}$ and $PRMSE_{kx}$.

Haberman and Sinharay (2010) recast the CTT approach in a MIRT context and showed that the value:

$$PRMSE_{k\theta M} = 1 - \tau_{k\theta}^2 / \tau_{k0\theta}^2 , \tag{19.25}$$

is equal to the IRT marginal reliability (Adams, Wilson, & Wang, 1997) of the scores for dimension $k$ where $\tau_{k\theta}^2$ is the dimension-specific error variance and $\tau_{k0\theta}^2$ is the dimension-specific total variance. When the items associated with a given subscore are modeled unidimensionally, the corresponding $PRMSE_{k\theta U}$ can be computed.[11] When $PRMSE_{k\theta M}$ is less than $PRMSE_{k\theta U}$, using the MIRT subscore provides little added value relative to the unidimensional subscore. On the other hand, if $PRMSE_{k\theta M}$ is greater than $PRMSE_{k\theta U}$, the MIRT subscore provides more diagnostic information than the corresponding unidimensional subscore. In addition to comparing PRMSE values within a CTT or MIRT framework, Haberman and Sinharay (2010) showed that these values can be compared across frameworks. For instance, $PRMSE_{ks}$ and $PRMSE_{k\theta M}$ could be compared to determine if there is added value to using MIRT subscores as opposed to raw subscores. They also showed that the results obtained using the CTT and MIRT approaches are generally comparable. As a point of clarification, within an IRT framework, the Haberman and Sinharay approach does not currently consider the added value of MIRT subscores relative to an overall unidimensional score.

With respect to the practicality of this method, Sinharay (2010) conducted an extensive review of operational programs that report subscores and employed a simulation to identify the characteristics that generally must be satisfied in order to have any hope of meeting the Haberman (2008) criterion. In short, he found that subscores should be comprised of a minimum of 20 items with disattenuated correlations between the subscores that are lower than 0.80. This assumes that the subscores are distinct, although the criterion may be satisfied for augmented subscores when there are fewer items and/or the disattenuated correlation between the subscores is less than 0.85. In either case, when the criterion is satisfied, modeling and linking parallel forms multidimensionally should be defensible. On

11 Haberman and Sinharay (2010) also established $PRMSE_{Vk\theta M}$, which is based on the test characteristic curves for each dimension.

the other hand, if the criterion is not satisfied, one may be better off treating the forms as essentially unidimensional and linking the scores as such.

## Construct Shift

When multidimensional data are modeled unidimensionally using a latent variable model such as IRT, examinee abilities will be a linear weighted composite of abilities on the underlying dimensions (Reckase, 2009; Wang, 1986). This is akin to projecting scores from a multidimensional space onto a single linear composite within the space (Gower & Dijksterhuis, 2004). It can be shown that the weight of each dimension in the composite can be approximated by the eigenvector, $\boldsymbol{\omega}$, associated with the largest principal component of the matrix $\mathbf{a}'\mathbf{a}$ where $\mathbf{a}$ is a matrix of factor loadings or item slopes. Based on this, the unidimensional item and ability parameters for the 2PL/GPCM can be approximated as a weighted composite of the M2PL/MGPCM parameters:

$$\begin{aligned} a_i &= \boldsymbol{a}_i'\boldsymbol{\omega} \\ b_{ik} &= \frac{-d_{ik}}{\boldsymbol{a}_i'\boldsymbol{\omega}} \\ \theta_j &= \boldsymbol{\theta}_j'\boldsymbol{\omega}\,. \end{aligned} \tag{19.26}$$

The vector $\boldsymbol{\theta}_j'\boldsymbol{\omega}$ is commonly referred to as a *reference composite*.

Conceptually, the contribution of a given factor in a composite score is based on the extent to which the items load on each dimension. A factor with higher loadings should weigh more heavily in the composite; hence, it should not come as a surprise that the weights for the reference composite are determined solely by the factor loadings/item slope parameters. To illustrate the idea of a composite visually, consider two hypothetical tests

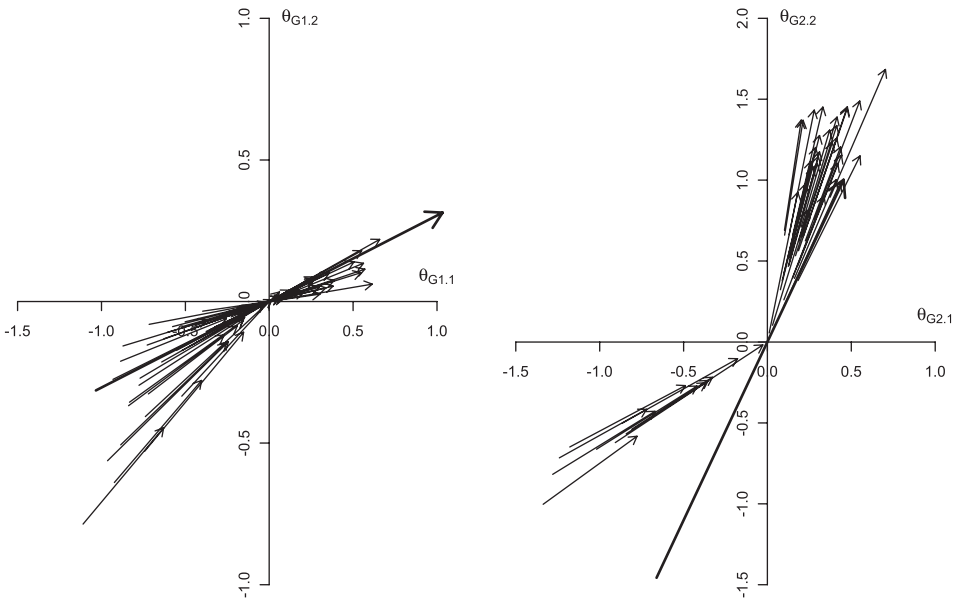


*Figure 19.2* Vector representation of items and corresponding reference composites.

that each measure two dimensions (see Figure 19.2).[12] The short arrows correspond to a vector representation of items (Reckase, 2009). The base of the arrow for item *i* corresponds to the multidimensional item difficulty (MDIF):

$$MDIF_i = \frac{-d_i}{\sqrt{\boldsymbol{a}_i' \boldsymbol{a}_i}}, \tag{19.27}$$

which characterizes the signed deviation from the origin. The length of the arrow corresponds to the multidimensional discrimination (MDISC):

$$MDISC_i = \sqrt{\boldsymbol{a}_i' \boldsymbol{a}_i} \ . \tag{19.28}$$

Longer arrows represent more discriminating items (items with more weight in the composite score). The direction of the arrow corresponds to the relative weight of each dimension. The angle of the arrow, relative to a given axis, *m*, is:

$$\alpha_{im} = arccos \frac{\boldsymbol{a}_{im}}{MDISC_i}, \tag{19.29}$$

where $\boldsymbol{a}_{im}$ is a specific element in $\boldsymbol{a}_i$. Arrows that are more closely aligned with a given axis measure that dimension more predominantly. The items on the first test (the left panel) load primarily on $\theta_1$, whereas the items on the second test load primarily on $\theta_2$. The reference composite for each test is characterized by the long, dark arrows spanning from approximately (–1.0, 0.25) to (1.1, 0.4) in the left panel and (–1.5, 0.75) to (0.5, 1.0) in the right panel.

Given that there are no substantive differences in the magnitude of loadings on these two dimensions (i.e., the lengths of the individual arrows are about the same for items loading on the two factors respectively), $\theta_1$ weighs more heavily in the composite for the first test simply because there are more items that measure that construct. Similarly, $\theta_2$ weighs more heavily in the composite for the second test. If we assume that the axes are aligned, there is an obvious change in the direction of the reference composite between the tests. This coincides with Martineau's (2004) conception of construct shift, which he characterizes as a change in the proportional representation of dimensions at different points along the composite scale. More formally, construct shift can be characterized by the angle between the two reference composites. The angle between the composites for tests *p* and *q* can be computed as:[13]

$$\alpha_{pq} = cos^{-1}\left[\frac{\boldsymbol{\omega}_p' \boldsymbol{\omega}_q'}{\sqrt{\boldsymbol{\omega}_p' \boldsymbol{\omega}_p}\sqrt{\boldsymbol{\omega}_q' \boldsymbol{\omega}_q}}\right]. \tag{19.30}$$

When reporting scores on a common scale there is a strong assumption that the measured construct(s) can be interpreted in the same way at all points along the scale. As such, the

12 The subscripts G1 and G2 in the axis labels denote group (test) 1 and 2 respectively.

13 Note that cos( $\alpha_{pq}$ ) = $r_{pq}$ where $r_{pq}$ can be interpreted as a correlation.

notion of a single construct (within and between tests) is sometimes interpreted as the presence of a single dominant factor with the possibility of several minor factors, for example, *essential unidimensionality* (Stout, 1987). This can be assessed by examining the extent to which the item covariances, conditional on examinee ability, are similar across scores. Conceptually, one can think about creating a reference composite for each ability level then comparing the direction of these composites. When the measurement direction of the composites differs substantively within and/or between tests—when there is construct shift—but the tests are modeled and linked unidimensionally, this raises questions about the comparability and interpretation of scores and any subsequent linking. One possible solution is to estimate the item parameters concurrently. Doing so will create a single reference composite that spans the multidimensional space for all of the tests being linked. This allows the scores to be compared on the same metric, but it does not resolve the problem of interpretation; what do scores on this composite scale mean substantively? If there is a clearer understanding of the underlying factors, it may be more valuable to model and link the tests multidimensionally.

## Common Item Requirements

Based on a consideration of the Haberman (2008) criterion for the added value of subscores and/or an examination of potential construct shift a decision may be made to establish a multidimensional scale for two or more test forms. If the data for the forms were collected based on a randomly equivalent groups design, the scores across dimensions can be scaled by fitting a single-group multidimensional model; however, if a nonequivalent groups common item design is used, the number of common items available to establish the linking must be considered. There is no minimum criterion for the number of common items that should be used to link tests, either unidimensionally or multidimensionally. Many researchers use the rule of 20 percent or 20 items (whichever is greater) as a minimum, whereas some testing companies require a minimum of 15 common items (cf., Sykes, 1997).

In general, more common items should produce a more stable linkage between tests, but in an era of short tests, the number of available common items may be limited. Studies conducted by Vale (1986) and Kim and Cohen (1998) suggest that linking error, in the unidimensional case, can be relatively small with as few as five dichotomous common items while other studies suggest that 10 to 15 common items are required for the error to be considered "acceptable" (Hanson & Béguin, 2002; Kim, 2006; Kim & Cohen, 2002). That is, if the tests as a whole are not very long, it may be possible to link the tests with fewer than 20 common items; however, potential bias in estimates of ability may be non-ignorable (Weeks, von Davier, & Yamamoto, 2014).

To address the issue of common item requirements for multidimensional linking, Yao (2010) conducted a simulation study to link tests with a between-item dimensional structure and found that when the dimension-specific scores are highly reliable, the linking may be adequate with as few as 10 common items overall and two to four common items per dimension; however, given that simulation was used, these requirements seem optimistically small. Weeks (2013) conducted a resampling study using empirical data from an assessment spanning multiple years. He found that for a between-item dimensional structure with a more restrictive criterion for the magnitude of error, the linking may be adequate with as few as 15 to 20 items per dimension, although a minimum of around 20 to 30 common items per dimension is preferable. Li and Lissitz (2000) conducted a simulation that considered multidimensional linking for items with complex structure (i.e., within-item dimensionality). They found that 15 to 20 common items overall may be sufficient to conduct the linking.

## Application

Dorans and Holland (2000) outlined a set of characteristics that are widely held as necessary for equating.[14] They include:

- Construct equivalence: The tests being equated should measure the same construct.
- Equal reliability: The test scores should have the same reliability.
- Score equity: Examinees' expected scores on the equated forms should be the same.[15]
- Symmetry of the equating function: The function used to transform scores from X to Y should be the inverse of the function used to transform scores from Y to X.
- Population invariance: The equating function used to transform scores from X to Y (or from Y to X) should be the same for all subpopulations.

In reality, none of these elements are likely to hold in the strictest sense (if they do, there is no need for equating), but they do provide a set of criteria by which to evaluate the comparability of equated scores. The goal of this section is to discuss these characteristics in the context of multidimensional test linking. To make the discussion more illustrative, empirical data from a large-scale mathematics assessment are presented.

## Design

Six scales are created for each of four linking scenarios (a total of 24 scales) using two years of data from grades 5 and 6. The linking scenarios include two equating conditions and two vertical scaling conditions.

Equating Conditions:

- Grade 5—Year 1 and Year 2
- Grade 6—Year 1 and Year 2

Vertical Scaling Conditions:

- Grade 5 and Grade 6—Year 1
- Grade 5 and Grade 6—Year 1 and Year 2

For each condition three models are applied to the data: a unidimensional model, a multidimensional model based on an exploratory factor structure with two dimensions, and a multidimensional model based on a confirmatory factor structure tied to the content standards. The scales for each model in each condition are created using both separate and concurrent calibrations. The details regarding the models and estimation methods are described later.

## Data

The data for the empirical illustration includes two years of item responses from a large-scale mathematics assessment in grades 5 and 6. There are around 58,000

14 These elements are based primarily on statements from Lord (1980).
15 Given the more practical test of equal reliability, for the present analysis, score equity is not evaluated.

examinees per grade, per year. Sixty-five percent of examinees are self-reported as white, 25 percent as Hispanic, 6 percent as black, and 3 percent as Asian/Pacific Islander. Thirty-four percent of examinees are on free or reduced lunch, and 9 percent are English language learners. The tests at each grade level include a combination of multiple-choice (MC) and constructed response (CR) items. Each year the grade 5 assessments include 54 MC items and 15 CR items. The grade 6 tests include 45 MC items and 15 CR items.

Each item on each of the grades 5 and 6 math assessments is associated with one of six content standards (based on the test blueprint): 1) Number Sense, 2) Algebra, Patterns, and Functions, 3) Statistics and Probability, 4) Geometry, 5) Measurement, and 6) Computational Techniques. In released reports, standards one and six and standards four and five are collapsed to produce four subscores: Std 1/6, Std 2, Std3, and Std 4/5. These standards are referred to here as "NUCO," "ALPF," "STPB", and "GEME" respectively (see Table 19.1).

The number of items associated with each of these standards does not change from year to year within a given grade; however, the proportional representation of each standard does change from grade to grade. In grade 5, nearly half of the items are NUCO items, but in grade 6 only 38 percent of the items measure this standard. ALPF and STPB are the least represented standards in grade 5 at 17 percent and 16 percent, respectively. In grade 6, there is a slight increase in the proportion of STPB items. There is also a notable increase in the proportion of grade 6 GEME items.

Table 19.2 shows the number of common items between tests for each dimension. In grade 5 there are a total of 33 common items (24%) between the Year 1 and Year 2 tests. In grade 6 there are 23 common items (19%) between the Year 1 and Year 2 tests, and between grades 5 and 6 there are 25 or 31 common items depending on whether the tests are linked using data from only Year 1 or if items are pooled across years within each grade prior to establishing the vertical scale. While the total number of common items in each case should be sufficient for unidimensional linking, the number of common items for most of the dimensions is quite small.

*Table 19.1* Content Standard Identification and Proportional Representation by Grade

| *Standard* | *Name* | *Description* | *Grade 5* | *Grade 6* |
|---|---|---|---|---|
| 1/6 | NUCO | Number Sense/Computational Techniques | 47% | 38% |
| 2 | ALPF | Algebra, Patterns, and Functions | 17% | 17% |
| 3 | STPB | Statistics and Probability | 16% | 18% |
| 4/5 | GEME | Geometry/Measurement | 20% | 27% |

*Table 19.2* Number of Common Items by Grade and Dimension

| | *NUCO* | *ALPF* | *STPB* | *GEME* | *Total* |
|---|---|---|---|---|---|
| Grade 5 (Y1–Y2) | 15 | 5 | 6 | 7 | 33 |
| Grade 6 (Y1–Y2) | 7 | 5 | 7 | 4 | 23 |
| Grades 5–6 (Y1) | 8 | 4 | 5 | 8 | 25 |
| Grades 5–6 (Y1/Y2) | 11 | 5 | 5 | 10 | 31 |

## Dimensionality

Considerable work has been done over the years to identify the dimensions underlying mathematical ability. Geary (2006), for example, provides evidence from studies in psychometrics, cognitive psychology, and behavioral genetics to suggest that there are two primary cognitive dimensions. The first dimension is associated with numerical facility (i.e., arithmetic computation, number relations) and spans all ability levels from infancy to adulthood. The second dimension is associated with mathematical reasoning (i.e., evaluation of quantitative relationships and drawing conclusions based on quantitative information) and spans most ability levels for school-aged children and beyond. His cited studies also indicate that there is a set of factors related to complex skills like algebraic problem solving and estimation that are only present at higher ability levels. At the same time, there is evidence to suggest that these cognitive dimensions are intertwined with content domains. That is, answering a question correctly may require simple recall, problem solving, or abstract reasoning (Hegarty & Kozhevnikov, 1999).

On the other hand, Abedi (1994) examined the dimensionality of the 1990 and 1992 National Assessment of Educational Progress (NAEP) mathematics assessments using a confirmatory factor analysis where the five reported subscores (number and operations, measurement, algebra, geometry, and statistics) served as the basis for the number of dimensions. Although he did not compare this structure to one based on exploratory factors, he found that this model fit the data very well. In other words, this model provides some evidence in support of a content-based factor structure. However, other studies suggest that when content areas are used as the basis for the factor structure, the factor correlations can be quite high, suggesting that mathematics assessments may be better fit by a unidimensional model (cf., von Davier & Sinharay, 2007). The results from these various studies, among others, suggest that if an exploratory approach is used to identify the dimensional structure of a math test, one is likely to identify a combination of content and process dimensions; however, these dimensions may not be consistently identified across grades. In contrast, a content-based structure—across grades—may be defensibly identified using a confirmatory analysis.

For this illustration, the data were examined using both exploratory and confirmatory multidimensional factor structures (as well as a unidimensional structure). With respect to the exploratory structure, an analysis of the dimensional structure of the mathematics data was conducted using a combination of parallel analysis (Horn, 1965) and a vector approach developed by Reckase, Martineau, and Kim (2000). Parallel analysis is an extension of principal components analysis that compares the magnitude of principal components in the empirical data to randomly generated data with the same number of variables. The vector approach, on the other hand, considers changes in the angles between item vectors—characterized by factor loadings or MIRT slopes—as the number of modeled dimensions increases. Parallel analysis has been shown to perform well when the underlying factors in the data are orthogonal; however, the number of dimensions may be underidentified if the factors are correlated. In this case, the vector approach may more accurately identify the number of underlying factors. The results from the parallel analysis and vector approach were consistent and suggest the presence of three to four underlying factors in each grade-level test (two of which are major factors); however, a comparison of model fit suggests that the data are better characterized by two dimensions. The exploratory models were each applied using two factors.

## Analyses

For all of the analyses, item parameters for the dichotomous items were estimated using the 2PL or M2PL; item parameters for the polytomous items were estimated using the GPCM or MGPCM. The data were modeled using three factor structures: 1) unidimensional, 2) two-factor within-item (exploratory), and 3) four-factor (confirmatory) simple structure based on the content standards. In each case the item parameters were estimated using IRTPRO (Cai, du Toit, & Thissen, 2012) via marginal maximum likelihood estimation.

For each factor structure, for each of the four linking scenarios described earlier, the item parameters for each test were first estimated separately for each test and then linked using a separate calibration approach (described later). The item parameters were then re-estimated concurrently such that the item parameters for the common items were constrained to be equal across groups. In linking scenario four where the items from Year 1 and Year 2 were pooled within each grade, the separate calibration approach proceeded by first estimating the item parameters within grade, across years concurrently and then estimating linking coefficients for the combined data. For the concurrent re-estimation, item parameters were estimated simultaneously using the data from both grades in both years. In all cases, the grade 5 and/or Year 1 test was treated as the reference scale and expected *a posteriori* (EAP) estimates of examinee ability were used.

With respect to the separate calibration approaches, in the unidimensional case, the Stocking-Lord method was implemented (with symmetry constraints) to maintain consistency with the linking of scales in operational use. In the multidimensional case, both for the exploratory and confirmatory factor structures, an oblique approach as well as orthogonal rotation with anisotropic scaling was used. For the oblique approach, symmetry constraints were applied. The estimation of linking coefficients for the separate calibrations were conducted using the plink package in R (Weeks, 2010).

## Construct Equivalence

As described earlier, when a test measuring multiple dimensions is modeled unidimensionally, examinee ability can be treated as a weighted composite of the underlying dimensions. Simply put, the scores should coincide with the largest principal component in the data. When the dimensions are highly correlated, the resulting scores can be treated as essentially unidimensional; on the other hand, if the dimensions are uncorrelated or only moderately correlated, the interpretation of the reference composite is more ambiguous. Setting aside the interpretation of composite scores, given that the reference composite characterizes the relative weight of each dimension within a test, a comparison of the angle between the composites on the tests—after they have been scaled multidimensionally—can be used to assess construct equivalence within a multidimensional framework. The requirement of construct equivalence should be satisfied if this angle is sufficiently small, or when the correlation is close to unity (there is no research that provides a criterion for this difference).

Table 19.3 presents the angles between the reference composites, expressed as correlations (values near one indicate construct equivalence), for the parallel forms in grades 5 and 6 (in Year 1 and 2 respectively) as well as the nonparallel forms between grades 5 and 6 in Year 1. The angles between the reference composite based on the multidimensional concurrent calibrations and the reference composites for each of the separately calibrated, linked scales are also presented. In all cases the correlations are high and none substantively differs from unity. This is evidence of little to no construct shift between the forms.

*Table 19.3* Angle Between Reference Composites (Expressed as Correlations)

| *Model* | *Comparison* | *Grade 5* | *Grade 6* | *Grade 5–6* |
|---|---|---|---|---|
| EFA | Y1 Sep—Y2 Sep | 0.999 | 0.996 | 0.994 |
| | Y1 Sep—Y1/Y2 Con | 0.950 | 0.951 | 0.951 |
| | Y2 Sep—Y1/Y2 Con | 0.951 | 0.955 | 0.957 |
| CFA | Y1 Sep—Y2 Sep | 0.995 | 0.998 | 0.946 |
| | Y1 Sep—Y1/Y2 Con | 0.998 | 0.999 | 0.998 |
| | Y2 Sep—Y1/Y2 Con | 0.995 | 0.999 | 0.946 |

Within grades one should expect this if the forms are designed to be parallel, whereas in the vertical scaling case where the test forms are not necessarily designed to be parallel, one might expect larger angles (lower correlations) between the composites. Regarding the differences between the reference composites from the separate calibration and the composites from the concurrent calibration, for the EFA comparisons, the composites based on the separate approach are more aligned than the separate and concurrent composites; however, the differences are still relatively small. In the CFA case, the composite for the concurrent calibration appears to be more closely aligned with the composite from the Year 1 separate calibration, as evidenced by the slightly lower correlation with the Year 1 separate calibration composite. Relative to the near perfect correlations for the within-grade conditions, particularly in the CFA case, there is some change—albeit small—in the angle for the between-grade linkage (e.g., 0.995 or 0.998 versus 0.946). This may be due in part to the change in representation of the NUCO and GEME standards from grades 5 to 6.

## Equal Reliability

Table 19.4 presents marginal reliabilities (Adams, Wilson, & Wang, 1997) for each factor under each dimensional structure, for each grade-by-year. The table also includes columns (N Inc) that identify the additional number of items required in a given grade-year for each dimension to have the same reliability as the more reliable test dimension in the pair; this is based on a reformulation of the Spearman-Brown equation (Lord & Novick, 1968). These values are provided primarily to aid in the interpretation of differences between the marginal reliabilities. For example, the reliabilities of the scores for the STPB dimension in grade 5 are 0.77 and 0.76 in Years 1 and 2, respectively. An additional two items associated with this dimension would need to be added to the Year 2 test in order for both tests to have reliabilities of 0.77.

The unidimensional scores in both grades, across years, are quite reliable under both the separate and concurrent approach. Most of the dimension-specific scales are moderately reliable, while some scales (e.g., the first factor for the EFA model) are not very reliable. With respect to equal reliability, the unidimensional model arguably satisfies this condition. Under the separate approach, several of the scales across years appear to satisfy the condition, while others do not seem to meet the condition. Perhaps one of the most notable differences is between the separate and concurrent calibration results. The marginal reliabilities for the concurrently calibrated scales are virtually indistinguishable between forms within a given grade, across years. This is further evidence in favor of concurrent calibration. The between-grade differences in reliability

*Table 19.4* Marginal Reliabilities

| | | *Grade 5* | | | *Grade 6* | | | *Grade 5–6 (Year 1)* |
|---|---|---|---|---|---|---|---|---|
| | | *Year 1* | *Year 2* | *N Inc* | *Year 1* | *Year 2* | *N Inc* | *N Inc* |
| Separate Calibration | UD | 0.93 | 0.93 | 5 | 0.93 | 0.94 | 6 | 6 |
| | EFA D1 | 0.55 | 0.52 | 7 | 0.75 | 0.62 | 20 | 54 |
| | EFA D2 | 0.82 | 0.81 | 4 | 0.83 | 0.80 | 7 | 7 |
| | NUCO | 0.89 | 0.88 | 2 | 0.85 | 0.86 | 1 | 20 |
| | ALPF | 0.77 | 0.81 | 14 | 0.72 | 0.76 | 10 | 19 |
| | STPB | 0.77 | 0.76 | 2 | 0.72 | 0.75 | 10 | 20 |
| | GEME | 0.74 | 0.78 | 13 | 0.78 | 0.81 | 10 | 0 |
| Concurrent Calibration | UD | 0.93 | 0.93 | 0 | 0.94 | 0.94 | 1 | 4 |
| | EFA D1 | 0.62 | 0.61 | 0 | 0.54 | 0.55 | 1 | 20 |
| | EFA D2 | 0.79 | 0.79 | 0 | 0.80 | 0.79 | 1 | 8 |
| | NUCO | 0.87 | 0.87 | 0 | 0.86 | 0.86 | 0 | 12 |
| | ALPF | 0.75 | 0.75 | 0 | 0.75 | 0.74 | 1 | 8 |
| | STPB | 0.71 | 0.71 | 1 | 0.75 | 0.75 | 1 | 1 |
| | GEME | 0.74 | 0.74 | 0 | 0.80 | 0.80 | 0 | 10 |

*Table 19.5* Observed and Disattenuated Correlations by Grade

| | | *UD* | *EFA D1* | *EFA D2* | *NUCO* | *ALPF* | *STPB* | *GEME* |
|---|---|---|---|---|---|---|---|---|
| Grade 5 | UD | | **0.86** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| | EFA D1 | 0.65 | | **0.36** | **0.82** | **0.80** | **0.87** | **0.95** |
| | EFA D2 | 0.90 | 0.25 | | **1.00** | **1.00** | **0.96** | **0.91** |
| | NUCO | 0.93 | 0.60 | 0.83 | | **0.97** | **0.93** | **0.93** |
| | ALPF | 0.86 | 0.54 | 0.78 | 0.78 | | **0.95** | **0.94** |
| | STPB | 0.83 | 0.58 | 0.72 | 0.73 | 0.69 | | **0.96** |
| | GEME | 0.83 | 0.64 | 0.69 | 0.74 | 0.69 | 0.70 | |
| Grade 6 | UD | | **0.76** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| | EFA D1 | 0.54 | | **0.22** | **0.90** | **0.66** | **0.78** | **0.56** |
| | EFA D2 | 0.91 | 0.14 | | **0.93** | **1.00** | **0.97** | **1.00** |
| | NUCO | 0.91 | 0.61 | 0.77 | | **0.94** | **0.93** | **0.92** |
| | ALPF | 0.85 | 0.42 | 0.80 | 0.75 | | **0.94** | **0.95** |
| | STPB | 0.85 | 0.50 | 0.75 | 0.74 | 0.70 | | **0.94** |
| | GEME | 0.88 | 0.37 | 0.85 | 0.77 | 0.73 | 0.73 | |

Note: Disattenuated correlations are located in the upper triangle.

are likely due, at least in part, to differences in the overall number of items in grades 5 and 6, 54 and 45 respectively. As such, it is not surprising that the number of items on the grade 6 test would need to be increased appreciably for the grade 6 scores to be as reliable as the grade 5 scores.

To further illustrate the impact of score reliability on the interpretation of dimension-specific scales, Table 19.5 presents the observed and disattenuated correlation between the ability estimates. The score reliabilities based on the concurrent calibration of items within each grade, across years, were used to disattenuate the correlations. The observed correlations between the CFA factors are positive and moderately high (0.69 to 0.78) while the correlations between the dimension-specific scores and the unidimensional scores are high (0.83 to 0.93). There is a low correlation between the factors for the EFA model, but moderate to high correlations with the unidimensional scores (0.54 to 0.91). However, after correcting for attenuation, most of the scales are indistinguishable from the unidimensional scale. Based on these results, there is strong evidence in favor of a unidimensional factor structure or, at best, a two-factor exploratory structure.

## Symmetry of the Equating Function

The notion of symmetry of the equating function in the context of IRT is premised on the idea that true scores are equivalent across tests. This is consistent with the assumption of parameter invariance. If this assumption holds, the regression of $\hat{\boldsymbol{\theta}}_{from}$ on $\hat{\boldsymbol{\theta}}_{to}$ should produce linking coefficients that are a near inverse to the regression of $\hat{\boldsymbol{\theta}}_{to}$ on $\hat{\boldsymbol{\theta}}_{from}$. In practice, the symmetry requirement is likely to be violated except in cases where the scores on each test are highly reliable. Given the expectation of symmetry, scaling functions are commonly specified in a manner that imposes this constraint. For instance, if the correlation is constrained to be unity in the regression of scores between the *to* and *from* tests, the same coefficients will be estimated regardless of the choice of the reference scale. Differences between coefficients estimated with and without this constraint could be used as evidence for the satisfaction of the symmetry requirement. Stated differently, if the linking coefficients obtained via unconstrained least squares are not appreciably different from the coefficients obtained using a method that imposes the symmetry constraint, this is strong evidence in favor of a symmetric relationship between the score scales.

Table 19.6 presents the diagonal of $\boldsymbol{T}$ when estimated using an unconstrained oblique Procrustes approach or the diagonal of $\boldsymbol{QS}$ in the orthogonal Procrustes approach. These values are reported as T1, T2, T3, T4. The values for the translation vector $\mathbf{m}$ (m1, m2, m3, m4) are also included. Note that the translation vector is not dependent on the rotation/dilation; thus, it is the same for both the oblique and orthogonal approaches. The unidimensional linking coefficients and estimates of the group means and standard deviations from the concurrent calibration are also reported. Although the reported diag($\boldsymbol{T}$) coefficients are not strictly dilation parameters, they do provide some evidence for the comparability of the transformations. In almost all cases, the dilation coefficients estimated under the oblique transformation are lower than the corresponding coefficients estimated under the orthogonal approach. This indicates that as more restrictive constraints are placed on the linking coefficients to obtain symmetry, adjustments to the variability of dimension-specific scores are likely to be higher. The differences in the dilation coefficients between the oblique and orthogonal approach are small to moderate (< 0.01 to 0.08) with a mean difference of 0.027 across all EFA transformations. This suggests that even without the symmetry constraint, the transformations are fairly symmetric. It is not possible to explicitly evaluate the symmetry of the transformations for the CFA cases because the linking coefficients for both transformations are identical (as expected). While not an evaluation of symmetry, the linking coefficients can be compared to the estimated group means and standard deviations from the concurrent calibration to assess the extent to which the concurrent and separate approaches provide the same adjustments to the

*Table 19.6* Dilation and Translation Coefficients and Concurrent Calibration Moments

| | | | | *T1* | *T2* | *T3* | *T4* | *m1* | *m2* | *m3* | *m4* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 5 | Separate Calibration | UD | | 1.06 | | | | 0.01 | | | |
| | | EFA | Oblique | 0.94 | 1.08 | | | –0.02 | 0.04 | | |
| | | | Orthogonal | 1.02 | 1.04 | | | | | | |
| | | CFA | Oblique | 1.03 | 1.45 | 1.04 | 1.02 | 0.19 | –0.12 | –0.21 | 0.10 |
| | | | Orthogonal | 1.03 | 1.45 | 1.04 | 1.02 | | | | |
| | Concurrent Calibration | UD | | 1.07 | | | | 0.09 | | | |
| | | EFA | | 1.17 | 1.01 | | | 0.90 | –0.44 | | |
| | | CFA | | 1.78 | 1.63 | 1.50 | 1.63 | 0.34 | 0.01 | 0.01 | 0.16 |
| Grade 6 | Separate Calibration | UD | | 0.95 | | | | –0.09 | | | |
| | | EFA | Oblique | 0.99 | 1.11 | | | 0.24 | –0.16 | | |
| | | | Orthogonal | 1.01 | 1.12 | | | | | | |
| | | CFA | Oblique | 1.04 | 0.99 | 1.00 | 1.01 | 0.10 | 0.01 | 0.30 | –0.10 |
| | | | Orthogonal | 1.04 | 0.99 | 1.00 | 1.01 | | | | |
| | Concurrent Calibration | UD | | 1.01 | | | | 0.09 | | | |
| | | EFA | | 1.37 | 1.12 | | | 0.30 | –0.05 | | |
| | | CFA | | 1.75 | 1.47 | 1.46 | 1.71 | 0.22 | 0.00 | 0.30 | 0.07 |
| Grades 5–6 | Separate Calibration (Year 1) | UD | | 0.99 | | | | 0.14 | | | |
| | | EFA | Oblique | 1.29 | 1.11 | | | 0.16 | 0.39 | | |
| | | | Orthogonal | 1.30 | 1.11 | | | | | | |
| | | CFA | Oblique | 1.37 | 1.34 | 1.39 | 1.20 | 0.65 | 0.58 | 0.40 | 0.30 |
| | | | Orthogonal | 1.37 | 1.34 | 1.39 | 1.20 | | | | |
| | Concurrent Calibration | UD G5 Y2 | | 1.08 | | | | 0.12 | | | |
| | | UD G6 Y1 | | 1.10 | | | | 0.34 | | | |
| | | UD G6 Y2 | | 1.13 | | | | 0.42 | | | |
| | | EFA G5 Y2 | | 1.03 | 1.15 | | | –0.45 | 1.02 | | |
| | | EFA G6 Y1 | | 1.05 | 1.35 | | | 0.33 | 0.18 | | |
| | | EFA G6 Y2 | | 1.28 | 1.42 | | | 0.58 | –0.02 | | |
| | | CFA G5 Y2 | | 1.77 | 1.65 | 1.52 | 1.66 | 0.28 | 0.14 | 0.00 | 0.17 |
| | | CFA G6 Y1 | | 1.80 | 1.53 | 1.62 | 1.66 | 0.64 | 0.44 | 0.43 | 0.41 |
| | | CFA G6 Y2 | | 1.83 | 1.62 | 1.61 | 1.76 | 0.82 | 0.36 | 0.68 | 0.38 |

underlying score distributions. In almost all cases, the concurrent calibration results in the largest adjustment to the variability of the scores; however, adjustments to the means are more ambiguous. There are no clear patterns in the magnitudes of translation coefficients and estimates of group means. This is likely due to the small numbers of common items for each dimension; the findings from Weeks (2013) suggest that a larger number of common items is needed to provide stable estimates of the translation coefficients, relative to the dilation coefficients.

## Population Invariance

For equated scores to be considered comparable, the function used to transform the scores should be the same for all subpopulations. This implicitly assumes that the item parameters for each test are population independent. In the multidimensional case this corresponds, in part, to the notion of measurement invariance. Meredith (1993) outlined four types of measurement invariance.

- Configural invariance: The pattern and relative magnitude of zero and nonzero loadings is maintained across groups.
- Weak invariance: The loadings for each item are equal across groups.
- Strong invariance: The loadings and intercepts for each item are equal across groups.
- Strict invariance: The loadings, intercepts, and residual variances for each item are equal across groups.

Multidimensional linking within an MIRT context assumes strong invariance (the notion of strict invariance is really only applicable in cases where an error term is included in the model; this is not the case with MIRT). Measurement invariance is commonly examined for different subpopulations on a single test, but in the case of multidimensional score linking, invariance must be evaluated across test forms. Under the concurrent approach where there is no secondary linking, the item parameters across subpopulations can be examined via a comparison of model fit based on imposed versus relaxed constraints on various item parameters across groups. With separate calibration, differences between transformed score estimates based on linking coefficients obtained from the overall sample versus subpopulation samples can be used to test the invariance assumption. Dorans and Holland (2000) introduced a root expected mean squared difference (REMSD) statistic to quantify subpopulation differences. Adapted for the multidimensional context, the formulation is:

$$REMSD_m = \frac{\sqrt{\sum_{h=1}^{H} w_h \mathbf{E}\{[e_{Th}(\theta_{Fm}) - e_T(\theta_{Fm})]^2\}}}{\sigma(\theta_{Tm})}, \tag{19.31}$$

where $e_T(\theta_{Fm})$ corresponds to the transformed *from* scale ability estimates for dimension $m$ on the *to* scale based on the linking coefficients obtained using the overall sample, $e_{Th}(\theta_{Fm})$ corresponds to the transformed scores for subgroup $h$ based on the linking coefficients obtained using data for the subgroup only, $w_h = N_h / N$ where $N$ is the number of examinees overall and $N_h$ is the number of examinees in the subgroup, $\sigma(\theta_{Tm})$ is the standard deviation of the *to* scale ability estimates, and $\mathbf{E}\{\}$ is the expected value.

Using the math data, item parameters were estimated separately and concurrently for each of the grade-level tests for each of the three-dimensional structures using the complete data and the full subset of data for males and females, respectively (in practice, the invariance assumption should be evaluated for other relevant subpopulations). In a given grade, in a given year there are approximately 29,000 males and 28,000 females. In the separate case, linking coefficients for each of these groups were estimated using the orthogonal Procrustes approach. To provide a reference for the magnitude of REMSDs under the separate approach, REMSDs were also computed using ability estimates from the concurrent calibration.

Table 19.7 presents the REMSDs for the various linkages for each dimensional structure. In general, the differences between the overall transformation and the subpopulation

*Table 19.7* Root Expected Mean Squared Deviations

| | | *UD* | *EFA D1* | *EFA D2* | *NUCO* | *ALPF* | *STPB* | *GEME* |
|---|---|---|---|---|---|---|---|---|
| Grade 5 | Separate | 0.016 | 0.111 | 0.171 | 0.007 | 0.019 | 0.041 | 0.009 |
| | Concurrent | 0.060 | 0.072 | 0.059 | 0.046 | 0.086 | 0.058 | 0.074 |
| Grade 6 | Separate | 0.003 | 0.310 | 0.337 | 0.005 | 0.003 | 0.011 | 0.018 |
| | Concurrent | 0.028 | 0.045 | 0.071 | 0.041 | 0.005 | 0.028 | 0.081 |
| Grades 5–6 | Separate | 0.013 | 0.159 | 0.312 | 0.040 | 0.039 | 0.050 | 0.013 |
| | Concurrent | 0.020 | 0.054 | 0.048 | 0.066 | 0.028 | 0.022 | 0.011 |

transformations are quite small, with the exception of the transformations for the EFA scales in the separate calibration case. This suggests that the exploratory factors may not have a consistent interpretation across subpopulations. In some instances the REMSDs for the separate case are slightly lower than the corresponding values in the concurrent case; however, the magnitude of the REMSDs are fairly comparable. In short, the transformations appear to be invariant for both the unidimensional and CFA models for both the separate and concurrent calibrations.

## Summary

As the use of multidimensional models becomes more prevalent, it is useful to consider the extent to which the requirements for test equating are satisfied in the multidimensional case. I considered four of the requirements outlined by Dorans and Holland (2000). In order to ensure that the substantive interpretation of equated scores is comparable, it is necessary for the tests to measure the same constructs, to the same degree. To test the requirement of construct equivalence, a comparison of the angle between the reference composites for the two tests can be used to determine if there is a notable change in the measurement emphasis between tests. In order to satisfy the requirement of construct equivalence, the representation of each dimension should be the same across tests; hence, the angle between the composites should be small.

The requirement of symmetry of the equating function is premised on the idea that true scores are equivalent across tests. Unless the scores are highly reliable, this requirement is unlikely to be satisfied; however, linking equations can be constrained to impose symmetry, or in the case of concurrent calibration, symmetry is imposed directly. The issue of symmetric transformations for multidimensional test score linking in the separate calibration case has been adequately addressed in the literature on multidimensional linking, but it is possible to obtain symmetric transformations by estimating an oblique transformation matrix via a two-sided Procrustes approach or through the use of an orthogonal rotation. Finally, the requirement of population invariance can be checked by examining the stability of estimated linking transformation for subpopulations of examinees using an REMSD statistic. Comparable transformations (low REMSDs) are suggestive of construct equivalence across subpopulations.

## Future Directions

The models and methods that form the basis for multidimensional test linking within an MIRT context have been in place for some time; however, the application of these methods in practice has been fairly limited. Much of the work up to this point has

centered on extending unidimensional linking methods (within a separate calibration framework), but more research is needed to address the adequacy of the methods for linking tests empirically on multiple dimensions. Any work related to multidimensional linking in a separate calibration framework should also be compared against corresponding concurrent calibrations. In addition, more work is needed to establish criteria for when it is defensible to establish a multidimensional scale relative to modeling the data unidimensionally as well as to refine the evaluation of the equating requirements in the multidimensional case.

## References


Abedi, J. (1994). *NAEP TRP Task 3e: Achievement dimensionality, section A*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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.

Baker, F.B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*(1), 87–96.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer-Verlag.

Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic.

Cai, L., du Toit, S., & Thissen, D. (2012). *IRTPRO: Flexible professional item response theory modeling for patient reported outcomes* [Computer Program].

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly*, *7*(4), 628–678.

Dorans, N.J., & Holland, P.W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306.

Dorans, N.J., Pommerich, M., & Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer-Verlag.

Embretson, S.E., & Wetzel, C.D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*(2), 175–193.

Geary, D.C. (2006). Development of mathematical understanding. In W. Damon (Ed.), *Handbook of child psychology* (6th ed., pp. 777–810). New York: John Wiley & Sons.

Gierl, M.J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT* (Research Report No. 2005–11). New York: College Board.

Gower, J.C., & Dijksterhuis, G.B. (2004). *Procrustes problems*. New York: Oxford University Press.

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

Haberman, S.J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229.

Haberman, S.J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Rep. No. RR-09–39). Princeton, NJ: Educational Testing Services.

Haberman, S.J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, (75)*2, 209–227.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. Japanese *Psychological Research, 22*(3), 144–149.

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*, *61*(1), 3–24.

Hegarty, M., & Kozhevnikov, M. (1999). Types of visual-spatial representations and mathematical problem solving. *Journal of Educational Psychology, 91*(4), 684–689.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Hull, C.L. (1922). The conversion of test scores into series which shall have any assigned mean and degree of dispersion. *Journal of Applied Psychology, 6*(4), 298–300.

Kelley, T.L. (1923). *Statistical method*. New York: Macmillan.

Kim, J. (2006). Using the distractor categories of multiple-choice items to improve IRT linking. *Journal of Educational Measurement*, *43*(3), 193–213.

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

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25–41.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53–76.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*(4), 219–226.

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

Kupermintz, H., Ennis, M.M., Hamilton, L.S., Talbert, J.E., & Snow, R.E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS:88 mathematics achievement. *American Educational Research Journal, 32*(3), 525–554.

Kupermintz, H., & Snow, R.E. (1997). Enhancing the validity and usefulness of largescale educational assessments: III. NELS:88 mathematics achievement to 12th grade. *American Educational Research Journal, 34*(1), 124–150.

Levine, R.E. (1955). *Equating the score scales of alternative forms administered to samples of different ability* (Research Bulletin 55–23). Princeton, NJ: Educational Testing Services.

Li, Y.H., & Lissitz, R.W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*(2), 115–138.

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

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179–193.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139–160.

Martineau, J.A. (2004). *The effects of construct shift on growth and accountability models*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Min, K. (2007). Evaluation of linking methods for multidimensional IRT calibrations. *Asia Pacific Education Review, 8*(1), 41–55.

Mulaik, S.A. (1972). *Foundations of factor analysis*. New York: McGraw-Hill.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Oshima, T.C., Davey, T., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*(4), 357–373.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M.D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401–412.

Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M.D., & Martineau, J.A. (2004). *The vertical scaling of science achievement tests* (Research report for the Center for Education and National Research Council).

Reckase, M.D., Martineau, J.A., & Kim, J.-P. (2000). *A vector approach to determining the number of dimensions needed to represent a set of variables*. Paper presented at the annual meeting of the Psychometric Society. Vancouver, Canada.

Schönemann, P.H., & Carroll, R.M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika, 35*(2), 245–255.

Sinharay, S. (2010). *When can subscores be expected to have added value? Results from operational and simulated data*. (ETS Research Rep. No. RR-10–16). Princeton, NJ: Educational Testing Services.

Skaggs, G., & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495–529.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617.

Sykes, R.C. (1997). *Guidelines for the selection of anchor items for mixed (or single) item format tests*. Monterey, CA: CTB/McGraw-Hill.

Thurstone, L.L. (1931). Multiple factor analysis. *Psychological Review, 38*(5), 406–427.

Thurstone, L.L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press.

Thurstone, L.L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.

Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*(4), 333–344.

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics, 32*(3), 233–251.

von Davier, M. & von Davier, A.A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*(3), 115–124.

Wainer, H., Vevea, J.L., Camacho, F., Reeve, B.B., Rosa, K., & Nelson, L. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, M.M. (1986). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT*. Paper presented at the Office of Naval Research Contractors Meeting.

Weeks, J.P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1–33.

Weeks, J.P. (2013). *Linking error in multidimensional vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education: San Francisco, CA.

Weeks, J. P., von Davier, M., & Yamamoto, K. (2014). Design considerations for the programme for international student assessment. In L. Rutkowski, M. von Davier, D. Rutkowski (Eds.), *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 259–275). Boca Raton, FL: Chapman Hall/CRC Press.

Yao, L. (2010). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35*(1), 48–66.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, *46*(2), 177–197.

Yao, L., & Schwarz, R.D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469–492.