Article

Exploring the Full-Information Bifactor Model in Vertical Scaling With Construct Shift

Applied Psychological Measurement
36(1) 3–20

DOI: 10.1177/0146621611432864
http://apm.sagepub.com

Ying Li[1] and Robert W. Lissitz[2]

Abstract

To address the lack of attention to construct shift in item response theory (IRT) vertical scaling, a multigroup, bifactor model was proposed to model the common dimension for all grades and the grade-specific dimensions. Bifactor model estimation accuracy was evaluated through a simulation study with manipulated factors of percentage of common items, sample size, and degree of construct shift. In addition, the unidimensional IRT (UIRT) model, which ignores construct shift, was also estimated to represent current practice. It was found that (a) bifactor models were well recovered overall, though the grade-specific dimensions were not as well recovered as the general dimension; (b) item discrimination parameter estimates were overestimated in UIRT models due to the effect of construct shift; (c) the person parameters of UIRT models were less accurately estimated than those of bifactor models; (d) group mean parameter estimates from UIRT models were less accurate than those of bifactor models; and (e) a large effect due to construct shift was found for the group mean parameter estimates of UIRT models. A real data analysis provided an illustration of how bifactor models can be applied to problems involving vertical scaling with construct shift. General procedures for testing practice were recommended and discussed.

Keywords

vertical scaling, full-information bifactor model, construct shift, item response theory (IRT), multidimensional IRT (MIRT), testlet model

Full-information bifactor models have been applied to empirical data from achievement tests to multiple-domain survey instruments along with unidimensional and other multidimensional models (Gibbons, Bock, Hedeker, Weiss, Segawa, & Bhaumik 2007; Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007). Among these applications, bifactor models were shown to be promising in terms of relative model fit over unidimensional and/or multidimensional models, and the discussion of bifactor model fit has been addressed at the item level by Li and Rupp (2011).

[1]American Institutes for Research, Washington, DC, USA
[2]University of Maryland, College Park, USA

Corresponding author:
Ying Li, American Institutes for Research, 1000 Thomas Jefferson Street NW, Washington, DC 20007, USA
Email: yingli@air.org

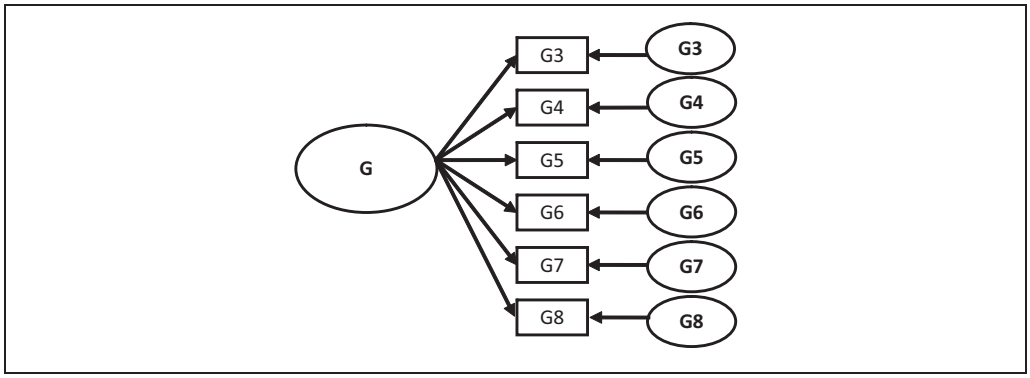

Figure 1. Illustration of a bifactor model for vertical scaling
Note: G = grade.

More recently, the bifactor model or its restricted form, the testlet model, has been successfully applied in testlet-based assessments to deal with psychometric issues such as vertical scaling (Li & Rijmen, 2009), differential item functioning (DIF) (Jeon, Rijmen, & Rabe-Hesketh, in press), and multigroup modeling (Cai, Yang, & Hansen, 2011; Jeon et al., in press); however, these applications of bifactor models are all limited to testlet-based assessments.

Bifactor models have potential for application to a much broader set of problems than testlet-based tests. The predominant reason for applying the bifactor model to broader contexts is its computational simplicity for estimation and its ease of interpretation. Because items in bifactor models can load on no more than one group-specific dimension in addition to the general dimension, no matter how many group-specific dimensions there are, the number of integrals for estimating any bifactor model is always two. Thus, the computational complexity of bifactor models is about the same as for two-dimensional MIRT models. In other words, the high-dimension bifactor models have a great advantage in computational simplicity over the high-dimension MIRT models. The relative simplicity of the bifactor model also facilitates interpretation and therefore, ease of application.

For another reason, the bifactor model structure (see Figure 1) aligns naturally with vertical scaling problems across grades. The general dimension in the bifactor model can be used to model the common vertical scale over grades; the group-specific dimensions can be used to model the grade-specific dimensions, or the constructs shifted from the general dimension. In addition, this modeling of vertical scaling is not limited to testlet-based exams or any limiting assumption that the test is unidimensional; instead, it is applicable to any set of tests that needs to be vertically scaled, no matter whether the tests are unidimensional or multidimensional at successive grade levels. In other words, the purpose of using the bifactor model for vertical scaling is to extract a common dimension for all grades and to set aside the residuals or grade-specific dimensions.

Item response theory (IRT) vertical scaling has two underlying assumptions: (a) unidimensionality of tests at each grade level and (b) test construct invariance across grades. Test unidimensionality means that test items measure a single latent trait at its targeted grade level; construct invariance across grades means that tests at different grade (or difficulty) levels maintain the same construct. Table 1 presents all the possible joint conditions of the two assumptions for IRT vertical scaling.

Shifts in constructs over grades have been demonstrated mathematically to significantly distort the results using vertical scales as outcomes (Martineau, 2004). Depending on the subject

Table 1. Joint Conditions of the Two Assumptions for Item Response Theory (IRT) Vertical Scaling

		Test invariance across grades	
		0 (*violated*)	1 (*satisfied*)
Test unidimensionality within grades	0 (*violated*)	(0, 0)	(0, 1)
	1 (*satisfied*)	(1, 0)	(1, 1)

matter, some tests tend to measure the same construct across grades better than others (e.g., Reckase & Martineau, 2004; Skaggs & Lissitz, 1988; Wang & Jiao, 2009). Absolute construct invariance is probably rarely, if ever, true.

A large number of studies (e.g., Hanson & Béguin, 2002; Kang & Petersen, 2009; Kim & Cohen, 2002; Meng, 2007; Tong & Kolen, 2007) have explored factors affecting IRT vertical scaling for cell (1, 1) in Table 1 when both assumptions hold, which is rarely true in applications. For cell (0, 1) where tests are multidimensional with construct invariance across grades, a few studies (Beguin & Hanson, 2001; Beguin, Hanson, & Glas, 2000; Patz & Yao, 2007; Simon, 2008) have applied multidimensional IRT models to vertical scaling. No studies have been found to model construct shift when construct invariance across grades is violated, that is, cell (1, 0) where tests are unidimensional within grades and cell (0, 0) where tests are multidimensional within grades in Table 1.

To address the lack of attention to construct shift in IRT vertical scaling, a bifactor model is proposed to estimate the common dimension for all grades and the grade-specific dimension for each grade. In addition, a unidimensional IRT (UIRT) model is estimated to represent the current practice in IRT vertical scaling.

This study has three objectives: (a) to propose and evaluate a bifactor model for IRT vertical scaling that can incorporate construct shift across grades while extracting a common scale, (b) to evaluate the robustness of the UIRT model in parameter recovery, and (c) to compare parameter estimates from the bifactor and UIRT models.

To achieve these objectives, specific research questions are raised:

Research Question 1: How well does the proposed bifactor model perform in recovering item and person parameters under various conditions of vertical scaling?

Research Question 2: How robust is the UIRT model in recovering item and person parameters under various conditions of the hypothesized true model for vertical scaling?

Research Question 3: How would the parameters estimated from the bifactor model and from the UIRT model differ under various conditions of the hypothesized true model for vertical scaling?

Method

The bifactor model's mathematical formulation and data collection designs for vertical scaling are briefly reviewed before the simulation study is introduced.

Bifactor Model

Gibbons and Hedeker (1992) generalized the work of Holzinger and Swineford (1937) to derive a bifactor model for dichotomously scored item response data. The model requires that (a) each

item have a nonzero loading on a general or common factor and only one nonzero loading on the group factors and (b) group factors are orthogonal to one another and to the general factor. For example, for a four-item test with two specific factors, the model might have the following factor pattern.

$$\begin{pmatrix} \alpha_{10} & \alpha_{11} & 0 \\ \alpha_{20} & \alpha_{21} & 0 \\ \alpha_{30} & 0 & \alpha_{32} \\ \alpha_{40} & 0 & \alpha_{42} \end{pmatrix},$$

where α_{ij} represents the loading of item i (i = 1, 2, 3, 4) on latent factor j (j = 0, 1, 2). Furthermore, the general factor across all grades is maximized in the bifactor model so that the common construct across grades can be maximally extracted while allowing variations at grade levels by modeling the grade-specific factors.

In the IRT framework, the probability of a correct response for an item i in the bifactor model can be modeled as

$$P(X_i = 1|\boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{is}\theta_s + d_i)]}, \quad (1)$$

where θ_0 represents the general factor or ability, θ_s(s = 1, 2, . . . , k) represents one of the k group-specific latent traits or abilities parameters that are mutually orthogonal and orthogonal to the general latent trait or ability parameter θ_0. Furthermore, a_{i0} and a_{is}(s = 1, 2, . . . , k) are item discrimination parameters for the general ability and one of the k group-specific abilities, respectively; as seen in the equation, for any item i, only one nonzero group-specific loading a_{is}(s = 1, 2, . . . , k) exists besides the general loading a_{i0}. Finally, d_i is a scalar parameter related to an overall multidimensional item difficulty.

The preceding general equation with θ_s representing one of the k group-specific abilities can be further written as a set of equations with group-specific abilities θ_1, θ_2, . . . , θ_k as

$$P(X_i = 1|\boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i1}\theta_1 + d_i)]}, \quad (2)$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i2}\theta_2 + d_i)]}, \quad (3)$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{ik}\theta_k + d_i)]}. \quad (4)$$

Data Collection Design for Vertical Scaling

The three data collection designs for vertical scaling illustrated by Kolen and Brennan (2004) are (a) common-item design, (b) equivalent groups design, and (c) scaling test design. Among the three designs, the common-item design is most popular and the easiest one to develop and implement; this is because greater overlap exists in subject curricula between adjacent grades for developing the common items. Specifically, the common-item design links adjacent grade assessments by including a set of common items in addition to the grade-level items. The graphical illustrations of the common-item design as well as the other two data collection designs can be found in Kolen and Brennan (2004, pp. 378-380).

Bifactor Model Specification for Data Collection Design

Multigroup bifactor models can be specified under all three data collection designs. Generally speaking, all the common items load only on the general factor, and grade-specific items load on corresponding grade-specific factors in addition to the general factor. To save space, only the bifactor model specification for the common-item design is illustrated here and model specifications for the other two designs can be found in Li (2011).

Under the common-item design, a set of common items is used for adjacent grades. For assessments from Grades 3 through 8, five sets of common items are needed; they are common items for Grades 3 and 4 (C34), common items for Grades 4 and 5 (C45), common items for Grades 5 and 6 (C56), common items for Grades 6 and 7 (C67), and common items for Grades 7 and 8 (C78). As seen in Figure 2, to specify a bifactor model, all items load not only on the general factor but also on the grade-specific factors.

Simulation Design

The research questions were approached through a simulation study. Three factors were fixed in the design: (a) the common-item data collection design was used, (b) the bifactor model was used as the true model for generating data, and (c) concurrent calibration was used.

The common-item design is the most often used data collection design for vertical scaling in practice. Many commercial and statewide vertically scaled testing programs apply the common-item design because it is relatively easy to create common items that are appropriate in terms of content and difficulty for adjacent grades.

Examinees' item response data were generated using bifactor models based on what was assumed about examinees' ability in vertically scaled assessments over grades. First, it was assumed that there is a single common scale (i.e., vertical scale) that captures examinees' growth over grades. Second, it was assumed that beyond this single general dimension, there are grade-specific dimensions or grade-related construct shifts from the single common dimension at the corresponding grade levels. Third, it was assumed that the single common dimension and the grade-specific dimensions are all orthogonal to one another, to allow unique explanations of the variances. From the individual perspective, two orthogonal latent ability scores were generated and estimated using the bifactor model: one general ability score that could be compared with students over all grades and one grade-specific ability score that could be compared across students at the same grade. Only three grade levels, conceptually labeled as Grades 3, 4, and 5, were considered in this study to represent the simplest scenario in vertical scaling. It is worth noting that factor analysis was conducted, and it confirmed the existence of primary and secondary dimensions in real vertically scaled assessment data, which provides evidence for the model used to generate the data.

Concurrent calibration requires only one computer run with response data for examinees at all grades to estimate item parameters simultaneously, whereas separate calibration requires one computer run for each grade, and linking methods are needed to place the estimates from multiple runs onto the same scale.

Concurrent calibration was selected over separate calibration in this study for several reasons. First, as pointed out by Kolen and Brennan (2004), when the IRT model holds, the concurrent calibration is expected to produce more stable results as it takes advantage of all the available information at once. In this study, because the true IRT model (i.e., the data generation model) was the bifactor model and the same model was used for calibration, the concurrent calibration should yield more reliable estimates. Second, concurrent calibration requires one computer run for all grades, but separate calibration requires multiple computer runs and linking methods be

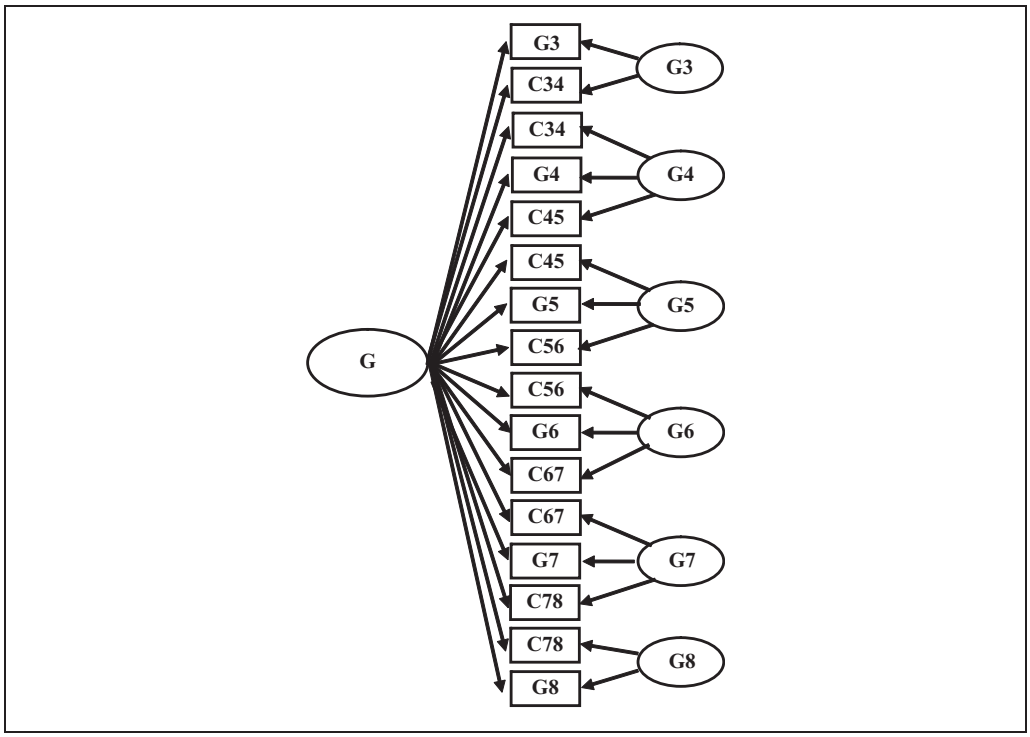

Figure 2. Bifactor model specification for common-item design
Note: G = grade; C34 = common items for Grades 3 and 4, C45 = common items for Grades 4 and 5, C56 = common items for Grades 5 and 6, C67 = Grades 6 and 7, C78 = common items for Grades 7 and 8.

applied, thus linking errors are unavoidable in separate calibration. Third, Simon (2008) compared concurrent calibration and separate calibration methods for simple structure MIRT models and concluded that concurrent calibration generally performed better because concurrent calibration benefited more from a larger sample size.

Three manipulated factors in the population bifactor data generation model were selected to investigate their effects on performance of the bifactor model estimation in vertical scaling. They were (a) sample size, (b) number or percentage of common items, and (c) variance of grade-specific factors.

As shown in Table 2, selecting three levels of sample size, three levels of number of common items, and three levels of grade-specific factors' variance created a study with $3 \times 3 \times 3$, or 27, fully crossed conditions. In all, 100 replications per condition were implemented.

In some studies addressing vertical scaling (Beguin et al., 2000; Beguin & Hanson, 2001; Smith, Finkelman, Nering, & Kim, 2008; Yon, 2006), sample size was fixed at 2,000 per grade; in other studies (Meng, 2007; Simon, 2008; Tong & Kolen, 2007), sample size was varied at three levels. In this study, to examine the effects of sample size on bifactor model vertical scaling, sample size was set at three levels: 1,000, 2,000, and 4,000 to represent relatively small, moderate, and large sample sizes.

Test length was fixed at 60 items in the study, and the percentage of common items varied. To satisfy the 20% common-items recommendation in all conditions (Kolen & Brennan, 2004), 20%, 30%, and 40% common items were used.

Table 2. Simulation Design

Factor	Level
Sample size (SS)	1,000; 2,000; and 4,000
% (*n*) of common items out of 60 (CI)	20% (12), 30% (18), and 40% (24)
Variance of grade-specific factors (VR)	0.25, 0.5, and 1

The degree of construct shift in bifactor model vertical scaling was measured by the variance of the grade-specific factors for the following reasons. In testlet models, or the constrained versions of bifactor models (Rijmen, 2010; Li, Bolt, & Fu, 2006), the variances of testlet factors are often manipulated to represent small, moderate, and large testlet effects; thus, in bifactor models, it is reasonable for the variances of group-specific factors to be manipulated to represent small, moderate, and large group-specific effects. Furthermore, in the context of bifactor model vertical scaling, the group-specific factors are grade-related factors, which represent the grade-specific constructs or the shifted constructs from the common factor. Note that when the variances of grade-specific factors are zeroes in the bifactor model, the bifactor model becomes a UIRT model, which has no construct shift.

To examine small, moderate, and large effects of grade-specific factors or shifted constructs in vertical scaling, the variances of the grade-specific factors were set at 0.25, 0.50, and 1.00, respectively. Only uniform effects of grade-specific factors were considered in the study; that is, the same magnitude of variance was used for all grade-specific factors in data generation.

Data Generation

Examinee latent ability was generated by a four-dimensional (the general dimension and the three grade-specific dimensions) multivariate normal distribution. Mathematically, this can be expressed as

$$\begin{pmatrix} \theta_0 & \theta_3 \\ \theta_0 & \theta_4 \\ \theta_0 & \theta_5 \end{pmatrix},$$

where θ_0 represents the general ability, θ_3, θ_4, and θ_5 represent grade-specific ability for Grade 3, 4, and 5 students, respectively. As shown, for any single examinee, there were two orthogonal latent abilities: the general ability and the grade-specific ability. Note that examinee latent ability was generated grade by grade, using multivariate distributions with grade-level means and unit standard deviations; Table 3 summarizes the detailed distributions of the latent trait parameter generation at each grade level.

Item discrimination parameters were set deliberately and repeatedly at 1.2, 1.4, 1.6, 1.8, 2.0, and 2.2 for the general dimension, and fixed at 1.7 for the grade-specific dimensions for the following reasons: First, 1.7 was chosen because it is the mean of the discrimination parameters for the general dimension (i.e., 1.2, 1.4, 1.6, 1.8, 2.0, and 2.2) to ensure that items discriminate well on general and specific dimensions; second, although it was possible to fix the discrimination parameters to different constant values, for simplicity's sake, the decision was made to fix all the discrimination values to a single constant.

The difficulty parameter b_i for item i was generated randomly from the normal distribution with means that were appropriate for the grade levels and with a fixed standard deviation of 1. Table 4 summarizes the generation of the difficulty parameter b_i for common and noncommon

Table 3. Latent Trait Parameter Generation

Grade level	General dimension	Grade-specific dimension		
	θ_0	θ_3	θ_4	θ_5
Grade 3	N(−0.5, 1)	N(0, 1)		
Grade 4	N(0, 1)		N(0, 1)	
Grade 5	N(+0.5, 1)			N(0, 1)

Note: N = normal distribution.

items. Once b_i was generated for tests at their grade levels, the scalar parameter d_i was computed by $d_i = -b_i\sqrt{a_{i0}^2 + a_{ij}^2}$ using the discrimination parameters a_{i0} and a_{ij} from the general dimension and one of the grade-specific dimensions (j = 3, 4, 5), respectively.

With both latent ability and item parameters generated, the item response function (i.e., Equation 1 or the set of Equations 2-4) was then applied to generate examinees' item response data grade by grade.

Identifications of Bifactor Model Estimation

To keep the bifactor models identified, for each of the uncorrelated latent dimensions, either the discrimination parameters (loadings) or the variance of the latent dimension needed to be fixed.

For the general dimension, as per convention, the variance of the general latent dimension was fixed to 1, and the discrimination parameters (loadings) were freely estimated in the study.

For the grade-specific dimensions (s = 1, 2, . . . , k), the discrimination parameters (s = 1, 2, . . . , k) (loadings) were fixed to the true parameter value 1.7, so that the variances of the grade-specific dimensions could be freely estimated. This decision was made because the intent of the study was to apply the bifactor model to vertical scaling with construct shifts; thus, being able to estimate the magnitudes of construct shifts or the variances of the grade-specific dimensions is essential.

Model Estimation

The side-by-side comparison of the bifactor estimation model and the UIRT estimation model is presented in Figure 3; note that the figure for the bifactor estimation model is the same as the bifactor data generation model.

Multigroup concurrent calibration was implemented. Students only answered their grade-level items and the common items from their adjacent grades; all other items were regarded as ''not reached.'' It is worth noting that concurrent calibration implies that all the items have unique sets of item parameter estimates (i.e., discrimination and difficulty-related scalar parameter estimates) including the common items, which were answered by students from two adjacent grades; in other words, the common items answered by multiple groups were restricted so that they would have unique item parameters in the multiple-group concurrent calibration.

The computer program IRTPRO[1] (Cai, Thissen, & du Toit, 2011), using marginal maximum-likelihood estimation (MML) with an EM algorithm, was used to estimate the models.

Evaluation Criteria

Bias, root mean square error (RMSE), and standard error (*SE*) were used to assess the accuracy of parameter estimates over the 100 replications at simulated conditions. Bias is the average difference between an estimate and the true parameter value over the replications; RMSE indicates

Table 4. Item Difficulty Parameter Generation

Type of item	Distribution of b_i parameters at Grade 3	Grade 4	Grade 5
Noncommon items	N(−0.5, 1)	N(0, 1)	N(+0.5, 1)
Common items	U(−1, 0.5)		
Common items		U(−0.5, 1)	

Note: N = normal distribution; U = uniform distribution.

the overall accuracy of parameter estimates; *SE* indicates the stability of parameter estimates. They were computed by averaging each of the values over all items or ability parameter estimates across replications:

$$\text{Bias}(\hat{\beta}) = \frac{\sum_{r=1}^{R} (\hat{\beta}_r - \beta)}{R}, \tag{5}$$

$$\text{RMSE}(\hat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R} (\hat{\beta}_r - \beta)^2}, \tag{6}$$

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R} \left(\hat{\beta}_r - \frac{\sum_{r=1}^{R} \hat{\beta}_r}{R} \right)^2}, \tag{7}$$

where β is the true ability or item parameter from data generated by the bifactor models, $\hat{\beta}_r$ is the estimated ability or item parameters at the rth replication ($r = 1, 2, \ldots, R$) from the bifactor estimation model, and R is the number of replications.

Results

Item Parameter Estimates

Aggregated bias, RMSE, and *SE* of item parameter estimates of bifactor models for each of the 27 simulated conditions are presented in Table 5. Generally speaking, item discrimination parameter estimates of the general dimension and item difficulty-related scalar parameter estimates were well recovered by the bifactor model estimations across the simulated factors. The results confirmed the expectation that with the increase of sample size, the estimation accuracy of the two item parameters would increase, and that with the increase of the degree of construct shift (i.e., the variance of grade-specific dimension), the estimation accuracy of the item discrimination parameters of the general dimension would decrease.

Person Parameter Estimates

Aggregated bias, RMSE, and *SE* of person parameter estimates (including the general and the grade-specific dimensions) of bifactor models for each of the 27 simulated conditions were computed, and comparisons were made for the general and the grade-specific dimension person estimates.

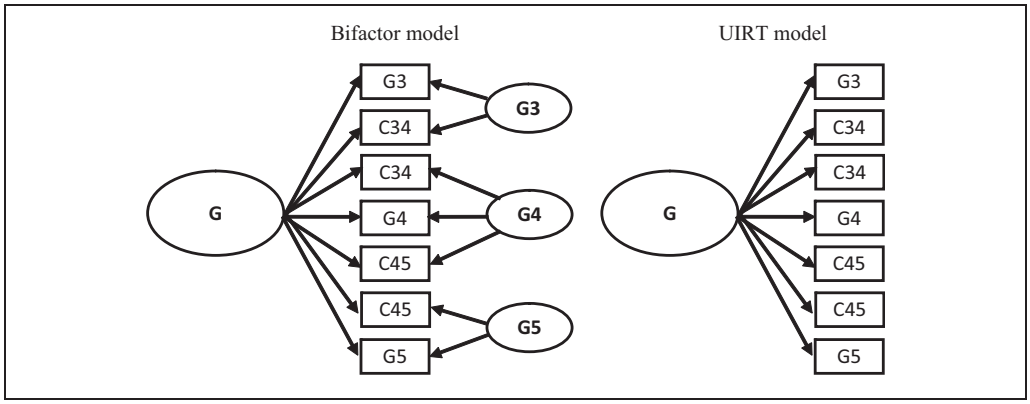

Figure 3. Bifactor versus UIRT estimation model
Note: G = grade; UIRT = unidimensional item response theory; C34 = common items for Grades 3 and 4; C45 = common items for Grades 4 and 5.

The results indicated that person parameter estimates of the general dimension were better recovered than that of the grade-specific dimensions when the degree of construct shift was small or moderate (i.e., the variance of grade-specific dimension was 0.25 or 0.50). Person parameter estimates of the general and grade-specific dimensions were about equally recovered when the degree of construct shift was large (i.e., the variance of the grade-specific dimension is 1.00). With the increase of sample size, the estimation accuracy of the person parameters of the general and grade-specific dimensions increased.

Group Parameter Estimates

Group mean estimates of the person parameters of the general dimension and variance estimates of the person parameters on the grade-specific dimensions are the two group parameter estimates in the bifactor model. To save space, only the aggregated bias and RMSE of the variance of the grade-specific dimension are presented in Figures 4 and 5 at the sample size of 4,000.

The results indicate that group mean parameters were well recovered across the simulated conditions; grade-specific variance parameters were also well recovered but were very slightly overestimated.

UIRT Models

Item discrimination parameters were greatly overestimated, whereas item difficulty-related scalar parameters were well recovered. For both item parameter estimates, it was found that with increases in sample size, estimation accuracy increased, and with increases in the degree of construct shift (i.e., variance of the grade-specific dimension), estimation accuracy decreased. Both person and group mean parameter estimates became less accurate as the degree of construct shift (i.e., variance of the grade-specific dimension) increased.

ANOVA Effects for the Simulated Factors

Three-way tests of between-subject effects (ANOVA) of bias, RMSE, and SE of all the parameter estimates in the bifactor and UIRT models for the three simulated factors were conducted. The ANOVA tests indicated that (a) sample size had small to moderate effects on item discrimination, difficultly, group mean ability, and grade-specific variance parameter estimates,

Table 5. Bias, RMSE, and *SE* of Item Parameter Estimate of Bifactor Models

			Discrimination parameter (a) estimate			Scalar parameter (b) estimate		
			VR			VR		
	SS	CI	0.25	0.50	1.00	0.25	0.50	1.00
BIAS	1,000	12	−0.05	−0.04	−0.05	0.02	0.06	−0.17
		18	−0.02	−0.01	−0.01	−0.08	−0.02	−0.03
		24	−0.06	−0.08	−0.06	0.01	0.09	0.03
	2,000	12	−0.04	−0.10	−0.08	0.06	−0.04	−0.01
		18	−0.05	−0.04	−0.13	−0.08	−0.06	0.01
		24	−0.08	−0.09	−0.07	0.08	0.06	0.03
	4,000	12	−0.07	−0.09	−0.11	−0.02	−0.04	−0.10
		18	−0.05	−0.09	−0.15	0.00	−0.04	−0.05
		24	−0.06	−0.10	−0.13	0.04	0.07	0.02
RMSE	1,000	12	0.22	0.22	0.25	0.25	0.23	0.29
		18	0.22	0.30	0.27	0.30	0.40	0.29
		24	0.20	0.24	0.27	0.21	0.27	0.25
	2,000	12	0.13	0.17	0.17	0.15	0.13	0.13
		18	0.14	0.15	0.22	0.18	0.18	0.17
		24	0.15	0.17	0.17	0.16	0.16	0.13
	4,000	12	0.11	0.13	0.15	0.09	0.10	0.14
		18	0.10	0.14	0.19	0.09	0.12	0.11
		24	0.10	0.14	0.16	0.09	0.11	0.09
SE	1,000	12	0.21	0.21	0.24	0.25	0.22	0.22
		18	0.21	0.29	0.27	0.28	0.39	0.28
		24	0.18	0.22	0.25	0.21	0.25	0.24
	2,000	12	0.12	0.13	0.15	0.13	0.12	0.13
		18	0.13	0.14	0.16	0.15	0.17	0.16
		24	0.12	0.13	0.15	0.14	0.14	0.12
	4,000	12	0.09	0.09	0.10	0.09	0.09	0.08
		18	0.09	0.10	0.10	0.09	0.11	0.09
		24	0.08	0.09	0.10	0.08	0.08	0.08

Note: RMSE = root mean square error; *SE* = standard error; SS = sample size; CI = common item; VR = variance of grade-specific factor.

as well as the stability of those parameter estimates; (b) the degree of construct shift (i.e., variance of the grade-specific dimension) had no or small effects on parameter estimates in bifactor models but had large effects on item discrimination and group mean ability parameter estimates in UIRT models; the degree of construct shift also had moderate to large effects on the stability of the parameter estimates in both models; (c) the percentage of common items resulted in a small amount of bias in difficulty parameter estimates in bifactor models but resulted in a large amount of bias in difficulty parameter estimates in UIRT models; also, the percentage of common items had small effects on the stability of the difficulty parameter estimates in bifactor models.

Comparisons of Bifactor and UIRT Model Estimates

Graphical comparisons of bias and RMSE for the person parameter estimates of the general dimension in the bifactor model and the person parameter estimates in the UIRT model are presented in Figures 6 and 7.

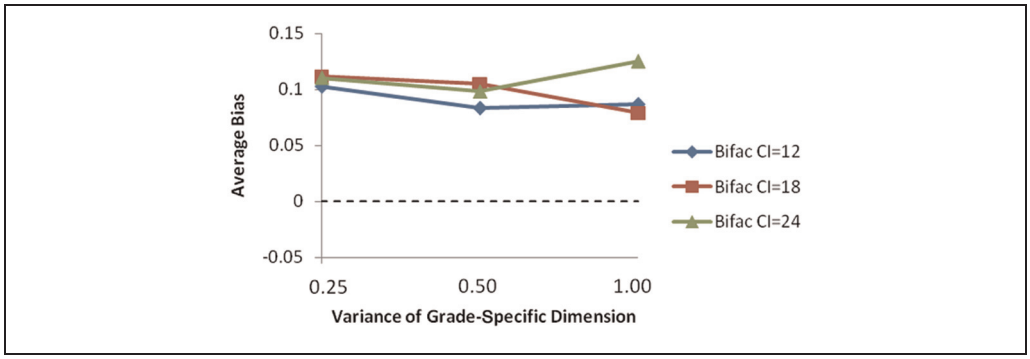

Figure 4. Mean bias of grade-specific variance parameter estimates at sample size of 4,000
Note: Bifac = bifactor models; CI = common items.

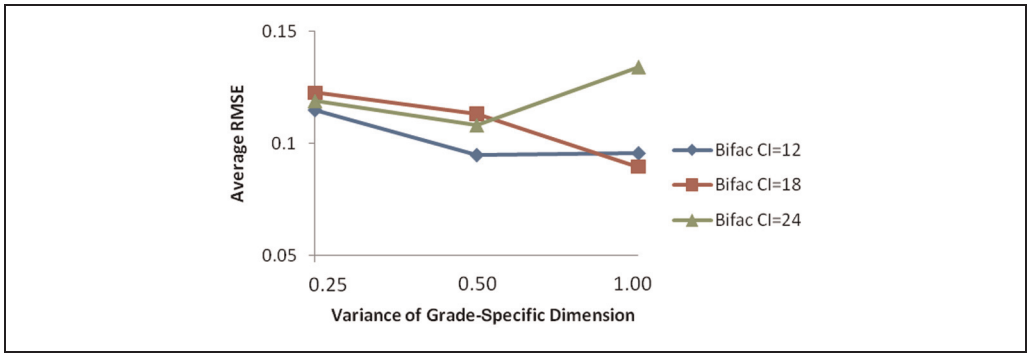

Figure 5. Mean RMSE of grade-specific variance parameter estimates at sample size of 4,000
Note: RMSE = root mean square error; Bifac = bifactor models; CI = common items.

Comparisons of all parameter estimates indicate that (a) item discrimination parameter estimates were overestimated in UIRT models due to the effect of construct shift (i.e., variance of the grade-specific dimension), which were underestimated to a smaller degree in bifactor models; (b) item difficulty parameters were well estimated in both UIRT and bifactor models, although bifactor model estimation resulted in somewhat smaller errors; and (c) person and group mean parameter estimates of UIRT models were always less accurate than that of bifactor models.

Real Data Application

Data

Empirical data from the 2006 fall Michigan mathematics assessments were obtained for Grades 3, 4, and 5. Michigan data were used because the tests had a common vertical scale, the common-item design was applied, and data from at least three consecutive years were available. Figure 8 shows the data collection design as well as the item distribution for the data. A total of 4,000 examinees were randomly selected from each grade for the data analysis.

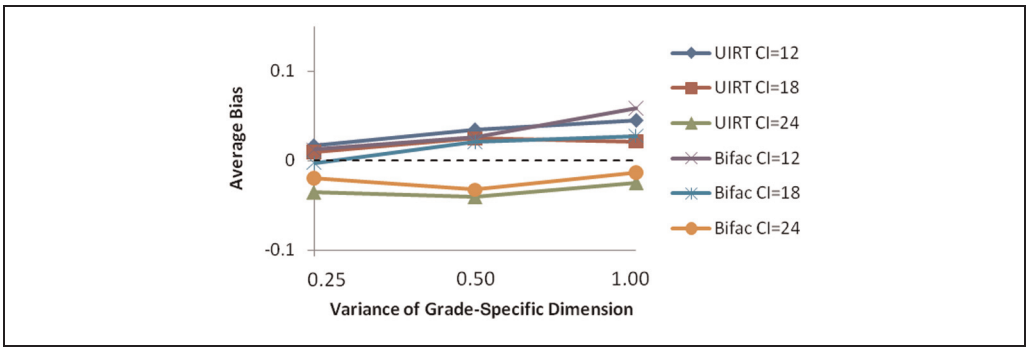

Figure 6. Mean bias of Person parameter estimates at sample size of 4,000
Note: Bifac = bifactor models; CI = common items.

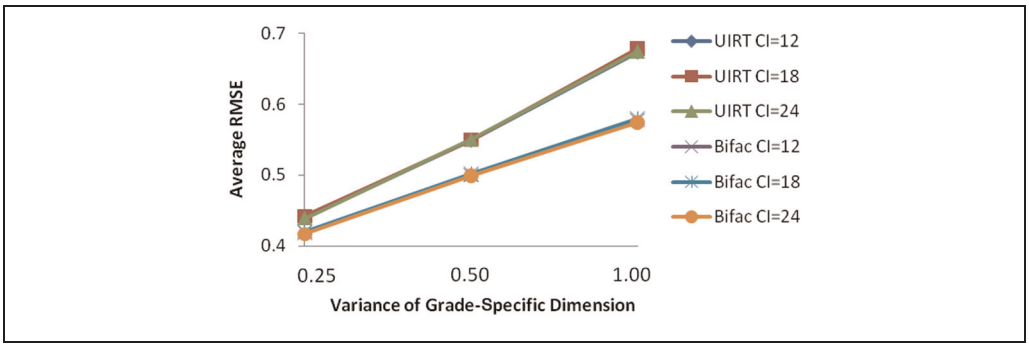

Figure 7. Mean RMSE of Person parameter estimates at sample size of 4,000
Note: RMSE = root mean square error; Bifac = bifactor models; CI = common items.

Research Questions

Three research questions were posed for the real data analysis:

Research Question 1: Which of the constrained bifactor models for vertical scaling with construct shift is the best-fitting model for the current data?

Research Question 2: What is the degree of construct shift in vertical scaling for these empirical data?

Research Question 3: How different are the general ability estimates from the best-fitting model that considers construct shift compared with the single ability estimates from the UIRT model, which ignores the construct shift?

Analysis

To achieve consistency with the simulation study, the same computer program, IRTPRO (Cai et al., 2011), was used for the real data analysis. Multiple bifactor models with different constraints were fit to the data to explore the degree of construct shift as well as to identify the best-fitting model. Akaike information criterion (AIC) and Bayesian information criterion (BIC) were calculated for model selection. Once the best-fitting bifactor model was determined,

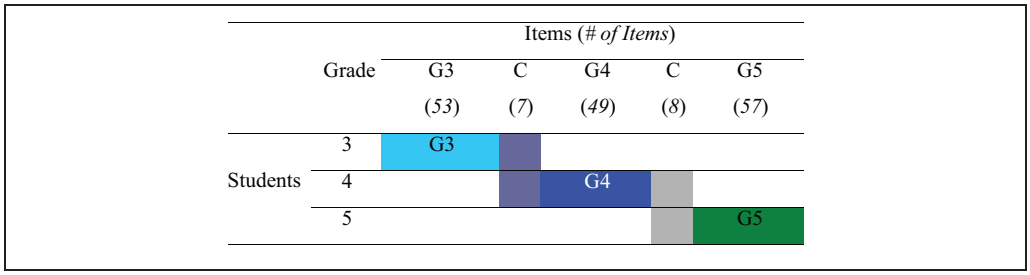

Figure 8. Data collection design and item distribution for the real data
Note: G = grade; C = common items.

a corresponding UIRT model was estimated to compare its parameter estimates with that of the bifactor model. Scatter plots and correlations were obtained for comparing person parameter estimates from the two models.

Procedures and Results

First, one should always assume construct shift across grades in vertical scaling. In other words, bifactor estimation models should always be considered for modeling vertical scaling with construct shift and for quantifying the degree of construct shift.

Next, the desire to quantify the degree of construct shift implies that the variances of the grade-specific dimensions in the bifactor model need to be estimated. Constrained bifactor models need to be specified to identify the estimation model and to make possible the estimation of variance of the grade-specific dimension.

Three constrained bifactor models were estimated for the current data. From least to most restrictive, they were (a) a bifactor model with fixed slopes (e.g., fixed to 1s) on the grade-specific dimensions, (b) a two-parameter testlet model, and (c) a Rasch (one parameter) testlet model. Note that for the general dimension, Grade 4 examinees were treated as the reference group and set to have a standard normal distribution. The standard deviations of the general dimension for the two nonreference groups were fixed to 1s to be consistent with the simulation study, though the standard deviations or variances of the general dimension for nonreference groups can be freely estimated. It is worth noting again that the common items answered by students from adjacent grades were restricted to have unique item parameter estimates in the multiple-group concurrent calibration. Table 6 reports the estimated variance of the grade-specific dimensions (i.e., the degree of construct shift), estimated group mean, as well as information criteria AIC and BIC for relative model fit.

As seen in Table 6, to find the best-fitting model for the first research question, the information criteria AIC and BIC values were calculated for the models; the smaller AIC and BIC values identified the better model-data fit; thus, the best-fitting model was the bifactor model with fixed slopes.

Using the estimated variances (0.21, 0.14, and 0.18) of the grade-specific dimensions from the best-fitting model, it was concluded, in answer to the second research question, that the degree of construct shift was small for the current data.

To approach the last research question on comparing person parameter estimates between the best-fitting model, which models construct shift, and the UIRT model, which ignores construct shift, a two-parameter UIRT model was also estimated for the current data. Table 7 reports the group estimates and information criteria for the two models selected.

Table 6. Group Estimates and Information Criteria for Constrained Bifactor Models

	Variance of the grade-specific dimension			Group mean on the general dimension			Information criteria	
Estimation model	G3	G4	G5	G3	G4	G5	AIC	BIC
Constrained bifactor	0.21	0.14	0.18	−0.61	0	0.19	**779,240**	**781,849**
2P testlet	0.33	0.54	1.06	−0.72	0	0.27	779,367	781,977
Rasch testlet	0.32	0.16	0.00	−0.63	0	0.22	789,191	790,514

Note: G = grade; 2P = 2-parameter. The 0.00 variance of the G3 dimension for the Rasch testlet model happened because the program encountered some difficulty with estimation and it seems that in one or more of the iterations, the variance went negative; in that case, the program sets the variances at the boundary of 0 and attempts to continue. The smallest values of information criteria are shown in bold face.

Table 7. Group Estimates and Information Criteria: Bifactor Versus UIRT Models

	Variance of the grade-specific dimension			Group mean on the general dimension			Information criteria	
Estimation model	G3	G4	G5	G3	G4	G5	AIC	BIC
Constrained bifactor	0.21	0.14	0.18	−0.61	0	0.19	**779,240**	**781,849**
2P UIRT	NA	NA	NA	−0.57	0	0.22	779,371	781,973

Note: G = grade; 2P = 2-parameter UIRT = unidimensional item response theory. The smaller values of information criteria are shown in bold face.

The AIC and BIC values in Table 7 indicate that the bifactor model with fixed slopes had a better model fit than the two-parameter UIRT model. It was found that the ability estimates from the two models were highly linearly related in the scatter plot and that the correlation of person parameter estimates from the two models was 0.983. Thus, the last research question was answered: The differences in person parameter estimates from the bifactor model, with fixed slopes, and the UIRT model are negligible, and the UIRT model provides simple and adequate results for vertical scaling for the current data.

The results of the real data analysis suggested that these data were closest to the simulated condition where the sample size was largest (e.g., 4,000), the number of common items was smallest (e.g., 12), and the degree of construct shift was smallest (e.g., variance of grade-specific dimension is 0.25). The findings of the real data analysis were consistent with that of the simulated condition.

Discussion

Simulated Factors

Sample size significantly affected parameter estimation and its stability; as sample size increased, parameter estimation accuracy and stability of parameter estimates increased. In the K-12 setting, sample size is usually very large, which favors parameter estimation. Variance of the grade-specific dimension (i.e., degree of construct shift) significantly affected the stability of parameter estimates; as the degree of construct shift increased, the stability of the general dimension estimates decreased and the stability of the grade-specific dimension estimates increased.

The number of common items had either no effect or a small effect in the simulation study. A quick review of the relevant literature suggested that under the UIRT linking, more common items are associated with smaller parameter estimation errors (e.g., Hanson & Béguin, 2002; Kim & Cohen, 2002; Meng, 2007), whereas under the MIRT linking, previous research (e.g., Simon, 2008) indicated that the percentage of common items had very small effects on parameter estimation accuracy, which is consistent with the finding of this study.

Implications for Testing Practice

Because the degree of construct shift significantly affected the stability of parameter estimates, the most important implication of this study for testing practice is to minimize the effect of construct shift by creating assessments that have substantial overlap between adjacent grades and to construct vertically scaled assessments across grades accordingly. As Yen (2007) pointed out, vertical scales that demonstrate growth over grades can be difficult to develop until the content standards, curricula, or test blueprints are designed to have hierarchical content strands with substantial overlap between grades.

After tests have been administered, practitioners should be cautious about construct shift, and exploratory analysis is needed to determine whether and when the bifactor model is an improvement to the use of the UIRT. The suggested procedures for determining the degree of construct shift in vertical scaling are as follows.

First, practitioners should always assume construct shift. Second, practitioners need to quantify the degree of construct shift by fitting different constrained bifactor models and to determine the best-fitting model using AIC and BIC. Third, the estimated variance of the grade-specific dimension will provide evidence for the degree of construct shift.

When the estimated degree of construct shift is small (i.e., ≤ 0.25), practitioners may want to apply the UIRT estimation model to see how person parameter estimates are different from the bifactor models. If the differences are large, the results from the best-fitting bifactor model should be used for vertical scaling with construct shift.

Limitations and Directions for Future Research

This study examined only one of the three data collection designs (i.e., common-item design) for vertical scaling. Investigations can be extended to the performance of the bifactor model vertical scaling for the equivalent groups design and the scaling test design.

In terms of item type, the current study considered tests with only dichotomously scored items. Future studies can be extended to polytomously scored items or even mixed item format tests. In terms of the bifactor item response function (Rijmen, 2010), this study considered a two-parameter (difficulty and discrimination) bifactor model; examination of a three-parameter (difficulty, discrimination, and guessing parameters) bifactor model or simplification to a one-parameter (difficulty parameter only) bifactor model can also be conducted.

Another limitation in this study lies in the bifactor model-data generation and estimation, where a single constant item discrimination or slope parameter value was generated for the grade-specific dimensions and the same constant was fixed in the bifactor model estimation. In practice, true parameter values are not known to practitioners and researchers. To deal with this issue, one may simply set the item discrimination parameters of the grade-specific dimension to unit values (1s) or to the mean of the discrimination parameter estimates of the general dimension when it is necessary to estimate the variance of the grade-specific dimension. Fitting several constrained bifactor models and calculating model fit indices will help determine the best-

fitting model as well as the degree of construct shift. The exploratory and confirmatory analyses of the real data provide good examples of this approach.

In terms of estimation method, this study applied the MML, implemented via the computer program IRTPRO (Cai et al., 2011). It would be interesting to compare different estimation methods. The available estimation methods include Bayesian estimation, using Markov chain Monte Carlo (MCMC) implemented via WINBUGS, and MML with an EM algorithm implemented via BNL (A Matlab toolbox for Bayesian networks with logistic regression nodes; Rijmen, 2006) and IRTPRO (Cai et al., 2011). Focus on parameter estimation accuracy as well as estimation time would be a worthwhile study.

Last but not least, as the grade-specific dimensions were not as well recovered as the general dimension in the bifactor model, future study might incorporate covariates (e.g., student background variables) to reduce the variance of the group-specific latent variables for bifactor models. Adding these covariates has been shown to change the results of value-added models (Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, & Resnick 2004) and perhaps they would do so in this context as well.

Acknowledgment

The authors thank Dr. Li Cai at the University of California at Los Angeles for making a beta version of the program available for this research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

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

Beguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the meeting of the National Council of Measurement in Education, New Orleans, IL.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO 2.1: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221-248.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., & Bhaumik, D. K. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41-54.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (in press). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*.

Kang, T., & Petersen, N. (2009). *Linking item parameters to a base scale* (ACT Research Report Series 2009-2). Iowa City, IA: ACT.

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

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

Li, F., & Rijmen, F. (2009, April). *A vertical linking design for periodic assessments and tests that consist of situated tasks*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Li, Y. (2011). *Exploring the full-information bifactor model in vertical scaling with construct shift* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21.

Li, Y., & Rupp, A. A. (2011). Performance of the S-X statistic for full-information bifactor models. *Educational and Psychological Measurement, 71*, 986-1005.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models*. ProQuest Information & Learning. Michigan State University, East Lansing, MI.

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling*. ProQuest Information & Learning. University of Iowa, Iowa City, IA.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253-272). New York, NY: Springer Science + Business Media.

Reckase, M. D., & Martineau, J. A. (2004). *The vertical scaling of science achievement tests* (Unpublished report). Michigan State University, East Lansing, MI.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16*, 19-31.

Rijmen, F. (2006). BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes [Computer software manual]. Amsterdam, Netherlands: Free University Medical Center.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT models. *Journal of Educational Measurement, 47*, 361-372.

Simon, M. K. (2008). *Comparison of concurrent and separate multidimensional IRT linking of item parameters*. ProQuest Information & Learning. University of Minnesota, Minneapolis, MN.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*, 69-82.

Smith, Z., Finkelman, M., Nering, M., & Kim, W. (2008, March). *Vertical scaling: A comparison of equating methods with unidimensional and multidimensional data*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*, 11-36.

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

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*, 760-777.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273-283). New York, NY: Springer.

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.