PSYCHOMETRIKA—VOL. 80, NO. 2, 535–559
JUNE 2015
DOI: 10.1007/s11336-014-9411-3

LORD–WINGERSKY ALGORITHM VERSION 2.0 FOR HIERARCHICAL ITEM FACTOR MODELS WITH APPLICATIONS IN TEST SCORING, SCALE ALIGNMENT, AND MODEL FIT TESTING

LI CAI

UNIVERSITY OF CALIFORNIA

Lord and Wingersky's (Appl Psychol Meas 8:453–461, 1984) recursive algorithm for creating summed score based likelihoods and posteriors has a proven track record in unidimensional item response theory (IRT) applications. Extending the recursive algorithm to handle multidimensionality is relatively simple, especially with fixed quadrature because the recursions can be defined on a grid formed by direct products of quadrature points. However, the increase in computational burden remains exponential in the number of dimensions, making the implementation of the recursive algorithm cumbersome for truly high-dimensional models. In this paper, a dimension reduction method that is specific to the Lord–Wingersky recursions is developed. This method can take advantage of the restrictions implied by hierarchical item factor models, e.g., the bifactor model, the testlet model, or the two-tier model, such that a version of the Lord–Wingersky recursive algorithm can operate on a dramatically reduced set of quadrature points. For instance, in a bifactor model, the dimension of integration is always equal to 2, regardless of the number of factors. The new algorithm not only provides an effective mechanism to produce summed score to IRT scaled score translation tables properly adjusted for residual dependence, but leads to new applications in test scoring, linking, and model fit checking as well. Simulated and empirical examples are used to illustrate the new applications.

Key words: multidimensional item response theory, bifactor model, testlet, linking, scale alignment, test equating, goodness-of-fit testing, summed score..

1. Introduction

The paper by Lord and Wingersky (1984) contains a terse description of a remarkably elegant recursive algorithm for computing summed score based likelihoods from the perspective of item response theory (IRT). According to Google Scholar, the paper has only been a moderate success in terms of citation counts (over 160 times as of this writing). However, the Lord–Wingersky algorithm motivated a number of important developments in educational and psychological measurement. For example, Thissen, Pommerich, Billeaud, and Williams (1995) extended the algorithm to test scoring with ordered polytomous IRT models. Thissen and Wainer (2001) presented a detailed account of related summed score based methods for test scoring using IRT, including methods for mixed-format tests involving a combination of multiple-choice (MC) and constructed response (CR) items. Orlando, Sherbourne, and Thissen (2000) applied the Lord–Wingersky algorithm to summed score based test linking. Chen and Thissen (1999) derived an item parameter calibration method based on summed scores. Orlando and Thissen (2000) proposed a solution to the item fit testing problem with a slight alteration of the original Lord–Wingersky algorithm.

Multidimensional IRT has flourished in recent years (see, e.g., Reckase, 2009). In particular, full-information item factor analysis (Bock, Gibbons, & Muraki, 1988) has become one of the central methodological pillars in educational and psychological measurement research (Wirth & Edwards, 2007). As IRT becomes adopted in new fields such as health-related patient reported

Correspondence should be sent to Li Cai, CRESST, University of California, Los Angeles, CA 90095-1521, USA. E-mail: lcai@ucla.edu

© 2014 The Psychometric Society

outcomes measurement (see Reeve et al., 2007), new item parameter estimation algorithms (e.g., Cai, 2010a; Edwards, 2010; Schilling & Bock, 2005) and flexible software implementations (e.g., Cai, 2013; Cai, Thissen, & Du Toit, 2011; Wu & Bentler, 2011) have emerged. The present paper is situated within this broad context.

One particular kind of confirmatory item factor analysis, full-information item bifactor analysis, has caught special attention among psychometric researchers (Gibbons & Hedeker, 1992). In an item bifactor model, all items load on a general dimension, and an item is permitted to load on at most one specific dimension. The specific dimensions are in essence group factors that account for residual dependence above and beyond the general dimension. The factor pattern in a bifactor analysis is an example of the *hierarchical* factor solution (Holzinger & Swineford, 1937; Schmid & Leiman, 1957).

The popularity of the item bifactor model has been, in no small part, due to Gibbons and Hedeker's (Gibbons & Hedeker, 1992) discovery of a dimension reduction method. With dimension reduction, maximum marginal likelihood estimation of item bifactor models requires at most two-dimensional numerical quadrature, irrespective of the number of factors in the model. Thus, truly high-dimensional confirmatory factor models may be fitted to item response data with reasonable numerical accuracy, computational stability, and most importantly, within a reasonable amount of time. Gibbons and Hedeker's (Gibbons & Hedeker, 1992) dimension reduction method did much to free item factor analysis from the "curse" of dimensionality.

The computational efficiency of the hierarchical item factor formulation prompted a flurry of recent activities in the technical literature (e.g., Gibbons et al., 2007; Jeon, Rijmen, & Rabe-Hesketh, 2013; Rijmen, Vansteelandt, & De Boeck, 2008; Rijmen, 2009), where new computational methods and extensions of the basic bifactor model are presented (see, e.g., Cai, 2010b; Cai, Yang, & Hansen, 2011). Within educational measurement, the closely related testlet response theory model (Wainer, Bradlow, & Wang, 2007) also garnered much attention. The testlet response theory model is a second-order item factor analysis model, but it is typically shown as a constrained version of item bifactor model (Glas, Wainer, & Bradlow, 2000; Li, Bolt, & Fu, 2006; Rijmen, 2010; Yung, McLeod, & Thissen, 1999).

Renewed interest in the hierarchical item factor model brings new methodological questions. As Reise (2012) noted, the bifactor model is appealing because it offers a convenient mechanism to accommodate nuisance multidimensionality without sacrificing the interpretability of the general dimension, which ultimately represents the target latent construct being measured, in contrast to other multidimensional IRT models (e.g., with multiple correlated latent variables). The existence of unequivocal general dimension(s) and the continued prevalence of summed scoring of assessment instruments imply that there is much theoretical and applied interest in a characterization of the relation between an observed summed score and its position on the latent general dimension(s), which calls for an extension of the classical Lord–Wingersky algorithm to the case of hierarchical item factor analysis models.

Even as one may extend the Lord–Wingersky algorithm to standard multidimensional IRT models using direct product quadrature rules, the computational complexity increases exponentially as more factors are added into the model. Therefore a different strategy is required—one that efficiently utilizes the restrictions implied by the hierarchical item factor analysis model to achieve dimension reduction analytically. The combination of Lord–Wingersky recursions with analytical dimension reduction results in what amounts to version 2.0 of the Lord–Wingersky algorithm. Its details will be the one of the foci of this paper.

With the availability of such an algorithm, a number of technical issues can be resolved. First, when multidimensional bifactor or testlet structures demonstrate superior fit to calibration data than the single-factor model, one can now construct summed score to IRT scaled score translation tables properly adjusted for residual dependence. Second, in terms of test linking, one can also achieve more than an extension of Orlando et al.'s (2000) summed score based method for linking

distinct groups. Thissen et al. (2011) *calibrated projection* method utilized two correlated general dimensions in a two-tier item factor model (Cai, 2010b) to produce the summed score to scaled score conversion table so that two closely related (yet not identical) instruments can be linked together with the method of projection. Third, the score combination methods for mixed-format tests described by Rosa, Swygert, Nelson, and Thissen (2001) can be obtained as a by-product of the Lord–Wingersky 2.0 algorithm, with no specialized computation required. Last but not the least, summed score computations can be useful for model fit checking. For instance, Orlando and Thissen's (2000) highly successful summed score based item fit statistic ($S-X^2$) can be extended to test item fit for bifactor models. The model-implied and observed summed score probabilities can also form diagnostic indices to check the ubiquitous latent variable normality assumption. The remainder of this paper will discuss each of the above applications in turn.

2. The Original Lord–Wingersky Algorithm

2.1. Summed Score Likelihoods

Let there be a total of $i = 1, \ldots I$ ordinal items. Let $T_i(k|\theta)$ be the ith item's traceline for category $k = 0, 1, \ldots, K_i - 1$. The summed scores range from 0 to $S = \sum_{i=1}^{I}(K_i - 1)$. From the perspective of IRT, the likelihood for the response pattern $\boldsymbol{u} = (u_1, \ldots u_I)$ can be expressed as

$$L(\boldsymbol{u}|\theta) = \prod_{i=1}^{I} T_i(u_i|\theta), \tag{1}$$

due to the assumption of independence of item responses conditional on the latent trait θ. Define $\|\boldsymbol{u}\| = \sum_{i=1}^{I} u_i$ as a notational shorthand for the summed score associated with response pattern $\boldsymbol{u}$. The likelihood for summed score $s = 0, \ldots, S$ is defined as

$$L(s|\theta) = \sum_{s=\|\boldsymbol{u}\|} L(\boldsymbol{u}|\theta) = \sum_{s=\|\boldsymbol{u}\|} \prod_{i=1}^{I} T_i(u_i|\theta), \tag{2}$$

where the summation in Eq. (2) is over all such response patterns that lead to a summed score equal to s. Given a population (prior) distribution $g(\theta)$, the unnormalized posterior for summed score s is

$$p(\theta|s) \propto L(s|\theta)\, g(\theta), \tag{3}$$

and the (marginal) probability for summed score s is

$$p(s) = \int L(s|\theta)\, g(\theta) \mathrm{d}\theta, \tag{4}$$

which implies that the normalized posterior of summed score s is

$$p(\theta|s) = \frac{L(s|\theta)\, g(\theta)}{p(s)}. \tag{5}$$

Therefore, the posterior mean is

$$E(\theta|s) = \frac{1}{p(s)} \int \theta L(s|\theta)\, g(\theta) \mathrm{d}\theta, \tag{6}$$

and the posterior variance is

$$V(\theta|s) = E\left(\theta^2|s\right) - E^2(\theta|s) = \frac{1}{p(s)}\int \theta^2 L(s|\theta)\, g(\theta)\mathrm{d}\theta - E^2(\theta|s). \tag{7}$$

The posterior mean and the square root of the posterior variance may be taken as the point estimate and the standard error of measurement for θ. The marginal probability, posterior mean, and posterior variance for the summed scores are key estimands that the IRT model can generate as long as the categories are ordered to allow for an approximate monotonic relationship between summed scores and scaled scores.

2.2. Dichotomous Item Responses

It is more convenient to introduce the Lord–Wingersky algorithm for the case of dichotomously scored items. The extension to polytomous data is straightforward (as shown in Section 2.5). For now, all K_i's are taken to be identically equal to 2. In this case, the maximum summed score S is equal to the number of items I. The definition in Eq. (2) requires evaluating all 2^I response pattern likelihoods, which becomes computationally intractable when I is large. On the other hand, Lord and Wingersky's (1984) algorithm builds the summed score likelihoods recursively, one item at a time. Let $L_i(s|\theta)$ denote the likelihood for summed score s, after item i has been added into the computation.

The algorithm starts by initializing the summed score likelihoods from item 1. As such, there are two possibilities $L_1(0|\theta) = T_1(0|\theta)$ and $L_1(1|\theta) = T_1(1|\theta)$ at the end of Step 1. Next, the second item is added. Note that at the end of the second step there will be three summed scores. The likelihood for summed score 0 is $L_2(0|\theta) = L_1(0|\theta)T_2(0|\theta)$. The likelihood for summed score 1 is a combination of two distinct possibilities: $L_2(1|\theta) = L_1(1|\theta)\,T_2(0|\theta) + L_1(0|\theta)\,T_2(1|\theta)$. The likelihood for summed score 2 is $L_2(2|\theta) = L_1(1|\theta)T_2(1|\theta)$. Then, in Step 3, item 3 is added. The likelihood for summed score 0 is $L_3(0|\theta) = L_2(0|\theta)\,T_3(0|\theta)$. The likelihood for summed score 1 is: $L_3(1|\theta) = L_2(1|\theta)\,T_3(0|\theta) + L_2(0|\theta)\,T_3(1|\theta)$. The likelihood for summed score 2 is: $L_3(2|\theta) = L_2(2|\theta)\,T_3(0|\theta) + L_2(1|\theta)\,T_3(1|\theta)$. Finally, the likelihood for summed score 3 is $L_3(3|\theta) = L_2(2|\theta)T_3(1|\theta)$. More generally, after initialization at item 1, in Step i of the recursive algorithm, item $i = 2, \ldots, I$ is added into the existing summed score likelihoods according to the following rules:

$$\begin{aligned} L_i(0|\theta) &= L_{i-1}(0|\theta)\,T_i(0|\theta), \\ L_i(s|\theta) &= L_{i-1}(s|\theta)\,T_i(0|\theta) + L_{i-1}(s-1|\theta)\,T_i(1|\theta), \quad \text{for } s = 1, \ldots i-1, \\ \text{and } L_i(i|\theta) &= L_{i-1}(i-1|\theta)\,T_i(1|\theta). \end{aligned} \tag{8}$$

The recursion is repeated until all I items have been added. At the end of the recursions, each accumulated $L_I(s|\theta)$ will be equal to the summed score likelihood $L(s|\theta)$ defined earlier in Eq. (2). As one can see, the recursive algorithm does not require explicitly enumerating all 2^I response pattern likelihoods.

In practice, because the integrals in Eqs. (4), (6), and (7) cannot be solved analytically, it is necessary to evaluate the summed score likelihoods over a set of quadrature points so that numerical summaries of the posterior can be computed. For instance, the marginal probability can be approximated to arbitrary precision using a Q-point rule:

$$p(s) = \int L(s|\theta)\, g(\theta)\mathrm{d}\theta \approx \sum_{q=1}^{Q} L\left(s|X_q\right) W(X_q), \tag{9}$$

TABLE 1.
Ordinates of item response functions and quadrature weights evaluated at five rectangular quadrature points for the three hypothetical items in the example.

θ	−2	−1	0	1	2
$T_1(1\|\theta)$	.032	.100	.269	.550	.802
$T_2(1\|\theta)$	.010	.232	.450	.690	.858
$T_3(1\|\theta)$	.269	.450	.646	.802	.900
$T_1(0\|\theta)$	.968	.900	.731	.450	.198
$T_2(0\|\theta)$	.900	.769	.550	.310	.142
$T_3(0\|\theta)$	.731	.550	.354	.198	.100
$W(\theta)$	.054	.244	.403	.244	.054

where X_q is a quadrature node and $W(X_q)$ is the corresponding quadrature weight. Gauss–Hermite quadrature is used extensively in the literature because the prior distribution of θ is typically assumed to be Gaussian. However, for simplicity, rectangular quadrature may be used, where $W(X_q)$ is a set of normalized ordinates of the prior density, i.e., $W\left(X_q\right) = g(X_q)/\sum_{q=1}^{Q} g(X_q)$, and the quadrature nodes are chosen to represent a sufficiently fine grid over an interval that captures most of the probability mass of the posterior, e.g., from −4 to +4, for a standard Gaussian prior.

2.3. *An Illustrative Example*

It is instructive to consider a simple test with three dichotomous items. The item tracelines are characterized by the two-parameter logistic model:

$$T_i\,(1|\theta) = \frac{1}{1 + \exp[-(c_i + a_i\theta)]}, \tag{10}$$

for the correct/endorsement response ($k = 1$), where c_i and a_i are the item intercept and slope parameters. The incorrect/non-endorsement response ($k = 0$) has a traceline that is equal to $T_i\,(0|\theta) = 1.0 - T_i\,(1|\theta)$. The intercept parameters for the three items are −1.0, −0.2, and 0.6, respectively. The slope parameters are 1.2, 1.0, and 0.8, respectively.

Table 1 shows the values of the tracelines evaluated at five equally spaced quadrature points at θ levels −2, −1, 0, 1, and 2, as well as the corresponding quadrature weights at each point. The quadrature weights are normalized ordinates of a standard Gaussian prior density for θ. Based on the item tracelines and weights in Table 1, one can apply the Lord–Wingersky algorithm to recursively accumulate the four summed score likelihoods (0, 1, 2, 3) for the three dichotomously scored items. Table 2 shows the recursive computations in some detail. As one can see, after the initializations in Step 1, the recursive algorithm follows Eq. (8) until all items have been added. The set of four summed score likelihoods at the end of Step 3 are represented numerically at the specified quadrature points. Of course, in practice, many more quadrature points are used for better precision. Table 2 merely serves as an illustration similar to Thissen and Wainer's (2001) Table 3.8 (p. 124).

With the quadrature weights in Table 1 and the summed score likelihoods in Table 2, one may directly compute the unnormalized summed score posteriors according to Eq. (3) by multiplying the summed score likelihood $L\,(s|\theta)$ with the prior weight $W\,(\theta)$ at each of the chosen quadrature points. Table 3 shows the posterior computations in detail. The unnormalized summed score posteriors are found by multiplying (point-by-point) the values of the summed score likelihoods

TABLE 2.
Accumulating the summed score likelihoods at five rectangular quadrature points for the three hypothetical items with Lord–Wingersky algorithm.

Summed score likelihoods	θ	-2	-1	0	1	2
Step 1: initialize summed score likelihoods by adding Item 1						
$L_1(0\|\theta) =$	$T_1(0\|\theta)$	.968	.900	.731	.450	.198
$L_1(1\|\theta) =$	$T_1(1\|\theta)$	.032	.100	.269	.550	.802
Step 2: add item 2 to existing summed score likelihoods						
$L_2(0\|\theta) =$	$L_1(0\|\theta)T_2(0\|\theta)$	.871	.692	.402	.140	.028
$L_2(1\|\theta) =$	$L_1(1\|\theta)T_2(0\|\theta) + L_1(0\|\theta)T_2(1\|\theta)$	.126	.285	.477	.481	.284
$L_2(2\|\theta) =$	$L_1(1\|\theta)T_2(1\|\theta)$	.003	.023	.121	.379	.688
Step 3: add Item 3 to existing summed score likelihoods						
$L(0\|\theta) = L_3(0\|\theta) =$	$L_2(0\|\theta)T_3(0\|\theta)$	.637	.380	.142	.028	.002
$L(1\|\theta) = L_3(1\|\theta) =$	$L_2(1\|\theta)T_3(0\|\theta) + L_2(0\|\theta)T_3(1\|\theta)$	.326	.468	.429	.207	.053
$L(2\|\theta) = L_3(2\|\theta) =$	$L_2(2\|\theta)T_3(0\|\theta) + L_2(1\|\theta)T_3(1\|\theta)$	.036	.141	.351	.461	.324
$L(3\|\theta) = L_3(3\|\theta) =$	$L_2(2\|\theta)T_3(1\|\theta)$	.001	.010	.078	.304	.620

TABLE 3.
Characterizing the summed score likelihoods and posteriors using the representation at five rectangular quadrature points for the three hypothetical items.

Quadrature	θ							
Weights at	-2	-1	0	1	2			
$W(\theta) =$	.054	.244	.403	.244	.054			
Summed score	θ							
Likelihoods $L(s\|\theta)$ at	-2	-1	0	1	2			
$L(0\|\theta) =$	.637	.380	.142	.028	.002			
$L(1\|\theta) =$	.326	.468	.429	.207	.053			
$L(2\|\theta) =$	.036	.141	.351	.461	.324			
$L(3\|\theta) =$	.001	.010	.078	.304	.620			
Unnormalized summed	θ					Posterior summaries		
Score posteriors $p(\theta\|s)$ at	-2	-1	0	1	2	$p(s)$	$E(\theta\|s)$	$V(\theta\|s)$
$p(\theta\|0) \propto L(0\|\theta)W(\theta) =$	.035	.093	.057	.007	.000	.19	$-.81$	.59
$p(\theta\|1) \propto L(1\|\theta)W(\theta) =$	.018	.114	.173	.051	.003	.36	$-.26$	.62
$p(\theta\|2) \propto L(2\|\theta)W(\theta) =$	.002	.034	.141	.113	.018	.31	.36	.61
$p(\theta\|3) \propto L(3\|\theta)W(\theta) =$	.000	.003	.031	.074	.034	.14	.98	.53

(the second panel) with the corresponding quadrature weights (the first panel). Summing over the quadrature representation of the unnormalized summed score posterior, as per Eq. (9), the marginal probabilities of the summed scores are shown in Table 3 under the column heading $p(s)$. These are the IRT model-implied probabilities for each of the summed scores. The posterior means $E(\theta|s)$ and posterior variances $V(\theta|s)$ are also presented in Table 3, essentially in the form of a summed score to IRT scaled score translation table. For instance, a summed score of 0 can be translated to an IRT scaled score of $-.85$ with standard error equal to the square root of .67. The probabilities can be used to construct percentile tables. Tables such as this facilitate the adoption of IRT scoring in practical situations.

2.4. Marginal Reliability of Scaled Scores

With the summed score to scaled score conversion table, a kind of marginal reliability coefficient can be computed for the scaled scores. Let $\bar{V}(\theta)$ denote the average error variance associated with θ. It may be obtained from the conversion table as a weighted sum

$$\bar{V}(\theta) = \sum_{s=0}^{S} V(\theta|s)\, p(s). \tag{11}$$

The marginal reliability of the scaled score conversions is defined as

$$\bar{\rho} = 1 - \frac{\bar{V}(\theta)}{\sigma^2(\theta)}, \tag{12}$$

where $\sigma^2(\theta)$ is the total (prior) variance of θ. From the results in Table 3, the average error variance is equal to 0.64. Since the latent trait θ has an assumed standard normal prior distribution, the total variance is 1.0. The marginal reliability of the scaled scores based on the summed scores is therefore equal to 0.36.

2.5. Polytomous Item Responses

Recall that $T_i(k|\theta)$ is the ith item's traceline for category $k = 0, 1, \ldots, K_i - 1$, and the number of categories ($K_i \geq 2$) may be different across items. Define $S_i = \sum_{j=1}^{i}(K_j - 1)$ as a notational shorthand for the maximum summed score after item i has been included. Clearly the maximum summed score is $S = S_I$.

The first step of the algorithm still involves the initialization of the K_1 summed score likelihoods at the category tracelines of item 1 so that $L_1(s|\theta) = T_1(s|\theta)$ for $s = 0, \ldots, S_1$. In Step $i = 2, \ldots, I$, the category tracelines of item i are added into the S_{i-1} available summed score likelihoods from the previous step, similar to the dichotomous case, but more complex book-keeping is required since the number of combinations leading up to the same summed score increases as the number of categories increases. For item i with K_i categories, and summed score $s = 0, \ldots, S_i$, the summed score likelihood can be written as

$$L_i(s|\theta) = \sum_{s_*=0}^{S_{i-1}} \sum_{k=0}^{K_i-1} L_{i-1}(s_*|\theta)\, T_i(k|\theta)\, \mathbf{1}_s(s_* + k), \tag{13}$$

where $\mathbf{1}_s(s_* + k)$ is an indicator function that takes on a value of 1 if and only if s is equal to $s_* + k$, and 0 otherwise. The summation in Eq. (13) is over the existing summed score likelihoods and K_i categories of item i, while preserving the restriction that the combination must lead to a summed score equal to s. Equation (13) reduces to the recursions in Eq. (8) when all items are dichotomous. After all I items have been added, $L_I(s|\theta)$ will become the desired summed score likelihood $L(s|\theta)$ for summed score $s = 0, \ldots, S$.

3. Lord–Wingersky Algorithm Version 2.0

3.1. A General Hierarchical Item Factor Model

Cai's (2010b) two-tier model represents a general hierarchical model that includes the standard (correlated-traits) multidimensional IRT model, item bifactor model, and testlet response

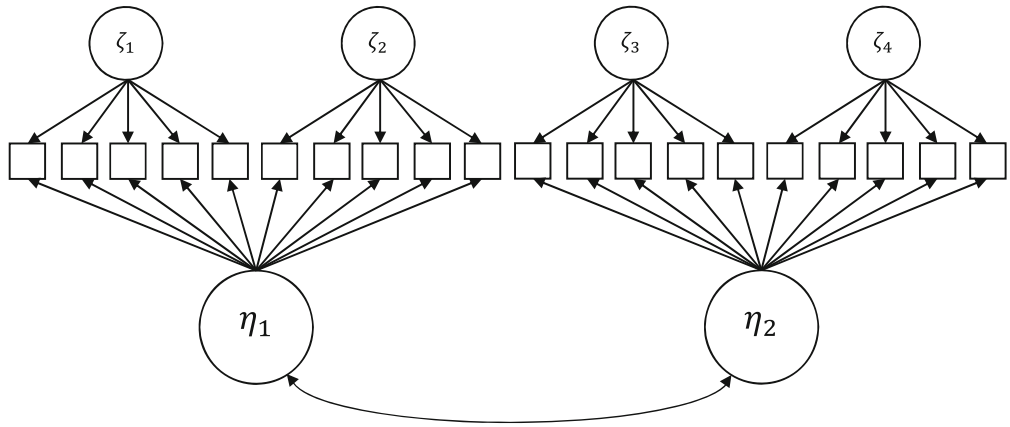

FIGURE 1.
Path diagram of a two-tier model with two correlated primary dimensions and four specific dimensions.

theory models as special cases. In this model, two kinds of latent variables are considered, primary and specific. This creates a partitioning of $\boldsymbol{\theta}$ into two mutually exclusive parts: $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\xi})$, where $\boldsymbol{\eta}$ is an M-dimensional vector of (potentially correlated) primary latent dimensions and $\boldsymbol{\xi}$ is an N-dimensional vector of (mutually orthogonal) specific latent dimensions that are orthogonal to the primary dimensions. In the two-tier model, an item is allowed to load on all M primary dimensions in any identified manner and at most one specific dimension. Using a path diagram, Figure 1 shows a hypothetical two-tier model with 20 items (the rectangles) that load on $M = 2$ primary dimensions that are correlated, as well as $N = 4$ specific dimensions. Obviously, a two-tier model with only one primary dimension becomes a bifactor or a testlet model.

Without loss of generality, let $T_i(k|\boldsymbol{\theta})$ be the ith item's traceline (or perhaps more properly referred to as trace-surface for multidimensional $\boldsymbol{\theta}$) for category k. In principle, the Lord–Wingersky algorithm can be defined on a set of quadrature points that are formed by direct products of unidimensional quadrature points. This leads to an exponentially increasing amount of computation in the number of latent dimensions. Fortunately, the two-tier formulation leads to a computational short cut that circumvents the integration problem. This is the main result of the paper.

3.2. General Approach

In the two-tier model, the item trace-surface $T_i(k|\boldsymbol{\theta})$ can be redefined as $T_i^n(k|\boldsymbol{\eta}, \boldsymbol{\xi}) = T_i^n(k|\boldsymbol{\eta}, \xi_n)$, for item i that loads on specific dimension n. The last equality comes from the fact that an item is permitted to load on at most one specific dimension, say, ξ_n in a two-tier model. If an item does not load on any specific dimension, it may be conveniently grouped with the first item cluster for the purposes of summed score computations and no generality is lost. Let there be I_n items that load on specific dimension ξ_n. As such, these I_n items form a testlet or item cluster that may be residually dependent after accounting for $\boldsymbol{\eta}$. For a two-tier model, the likelihood for response pattern $\boldsymbol{u}$ can be expressed as

$$L(\boldsymbol{u}|\boldsymbol{\theta}) = L(\boldsymbol{u}|\boldsymbol{\eta}, \boldsymbol{\xi}) = \prod_{n=1}^{N} \prod_{i=1}^{I_n} T_i^n\left(u_i^n|\boldsymbol{\eta}, \xi_n\right), \tag{14}$$

where u_i^n is the response to item i in item cluster n. Let $g_n(\xi_n)$ be the density function of the nth specific dimension. Integrating out the dependence on $\boldsymbol{\xi}$, the likelihood of $\boldsymbol{\eta}$ based on pattern $\boldsymbol{u}$ can be written as

$$\begin{aligned} L(\boldsymbol{u}|\boldsymbol{\eta}) &= \int \cdots \int \left[\prod_{n=1}^{N} \prod_{i=1}^{I_n} T_i^n \left(u_i^n | \boldsymbol{\eta}, \xi_n \right) \right] g_1(\xi_1) \cdots g_N(\xi_N) \mathrm{d}\xi_1 \cdots \mathrm{d}\xi_N \\ &= \prod_{n=1}^{N} \int \prod_{i=1}^{I_n} T_i^n \left(u_i^n | \boldsymbol{\eta}, \xi_n \right) g_n(\xi_n) \mathrm{d}\xi_n, \end{aligned} \tag{15}$$

where the second line in Eq. (15) have utilized the two-tier model assumption of the independence of the specific dimensions, thereby transforming the original N-fold multiple integral on the first line into a product of N onefold integrals. This is the same derivation as the dimension reduction procedure in maximum marginal likelihood item parameter estimation for two-tier or bifactor/testlet models (see, e.g., Cai, 2010b). Let

$$L^n(\boldsymbol{u}_n|\boldsymbol{\eta}) = \int \prod_{i=1}^{I_n} T_i^n \left(u_i^n | \boldsymbol{\eta}, \xi_n \right) g_n(\xi_n) \mathrm{d}\xi_n \tag{16}$$

denote the likelihood of $\boldsymbol{\eta}$ based on the subset of responses $\boldsymbol{u}_n = (u_1^n, \ldots, u_{I_n}^n)$ in the nth item cluster such that $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots \boldsymbol{u}_n, \ldots, \boldsymbol{u}_N)$. The likelihood of $\boldsymbol{\eta}$ for summed score s can be written as

$$L(s|\boldsymbol{\eta}) = \sum_{s=\|\boldsymbol{u}\|} L(\boldsymbol{u}|\boldsymbol{\eta}) = \sum_{s=\|\boldsymbol{u}\|} \prod_{n=1}^{N} L^n(\boldsymbol{u}_n|\boldsymbol{\eta}), \tag{17}$$

which is entirely analogous to Eq. (2). Integrating over $\boldsymbol{\eta}$, the marginal probability is $p(s) = \int L(s|\boldsymbol{\eta}) h(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\eta}$ (cf. Eq. 4), where $h(\boldsymbol{\eta})$ is the density of the primary dimensions, and the summed score posterior is $p(\boldsymbol{\eta}|s) = L(s|\boldsymbol{\eta}) / p(s)$ (cf. Eq. 5).

The dominating insight from Eq. (17) is that *conditional on the general dimension(s), the testlets or item clusters become the fungible units of model building and computation*, just as items are the fungible units in the standard Lord–Wingersky recursions. All that is required is an extra stage of recursions. In the first stage, for the nth item cluster, likelihoods for the within-cluster summed scores are accumulated over the latent variable space spanned by $(\boldsymbol{\eta}, \xi_n)$. This is standard Lord–Wingersky algorithm as applied to the items in cluster n on a set of direct product quadrature points spanning the space of $(\boldsymbol{\eta}, \xi_n)$. For each within-cluster summed score likelihood, the dependence on the specific dimension ξ_n is subsequently integrated out, leaving the within-cluster summed score likelihoods as functions of the general dimension(s) $\boldsymbol{\eta}$ alone. In the second stage, the N clusters are treated as N multiple-category items, and the within-cluster summed score likelihoods from the first stage are treated as if they are category tracelines defined on $\boldsymbol{\eta}$. Standard Lord–Wingerksy algorithm for polytomous IRT is applied to accumulate the final summed score likelihoods.

3.3. Details of the Lord–Wingersky 2.0 Algorithm

To avoid notational clutter, it would be convenient to introduce the new Lord–Wingersky algorithm for hierarchical item factor models using one of the simplest two-tier models, namely, the logistic item bifactor model for dichotomous responses. In this case, $T_i^n(k|\boldsymbol{\eta}, \xi_n)$ reduces

further to $T_i^n\,(k|\eta,\xi_n)$, and η represents the single general dimension. The IRT model for the correct/endorsement response can be written as

$$T_i^n\,(1|\boldsymbol{\theta}) = T_i^n\,(1|\eta,\xi_n) = \frac{1}{1+\exp\left[-\left(c_i + a_i^0\eta + a_i^n\xi_n\right)\right]}, \tag{18}$$

Note that there are two slope parameters per item in the bifactor model (cf. Eq. 10). The slope for the general dimension is a_i^0 and the slope for the nth specific dimension is a_i^n. The item intercept continues to be denoted as c_i.

With no loss of generality, consider the nth *item cluster*. The first stage of Lord–Wingersky algorithm 2.0 starts with the initialization of the within-cluster summed score likelihood: $L_1^n\,(0|\eta,\xi_n) = T_1^n\,(0|\eta,\xi_n)$ and $L_1^n\,(1|\eta,\xi_n) = T_1^n\,(1|\eta,\xi_n)$. Then, each of the remaining items within the cluster is added to the likelihoods according to the following set of recursions for $1 < i \leq I_n$ (cf. Eq. 8):

$$\begin{aligned}
L_i^n\,(0|\eta,\xi_n) &= L_{i-1}^n\,(0|\eta,\xi_n)\,T_i^n\,(0|\eta,\xi_n)\,,\\
L_i^n\,(s|\eta,\xi_n) &= L_{i-1}^n\,(s|\eta,\xi_n)\,T_i^n\,(0|\eta,\xi_n) + L_{i-1}^n\,(s-1|\eta,\xi_n)\,T_i^n\,(1|\eta,\xi_n)\,,\\
&\qquad \text{for } s = 1,\ldots i-1,\\
\text{and } L_i^n\,(i|\eta,\xi_n) &= L_{i-1}^n\,(i-1|\eta,\xi_n)\,T_i^n\,(1|\eta,\xi_n)\,.
\end{aligned} \tag{19}$$

At the end of the recursions the within-cluster summed score likelihoods will have been accumulated as $L_{I_n}^n\,(i|\eta,\xi_n) = L^n\,(s|\eta,\xi_n)$ for $s = 0,\ldots,r_n$, where $r_n = \sum_{i=1}^{I_n}(K_i - 1)$ is the maximum within-cluster summed score for item cluster n. Integrating out the dependence on ξ_n, the summed score likelihood as a function of η can be approximated with quadrature as

$$L^n\,(s|\eta) = \int L^n\,(s|\eta,\xi_n)\,g_n\,(\xi_n)\,\mathrm{d}\xi_n \approx \sum_{q=1}^{Q} L^n\left(s|\eta,X_q\right)W_n(X_q), \tag{20}$$

where X_q is a set of Q rectangular quadrature points with weights $W_n(X_q) = g_n(X_q)/\sum_{q=1}^{Q} g_n(X_q)$. At the end of the first stage, each of the N item clusters is characterized by a set of summed score likelihoods in terms of η.

In the second stage, $L^n\,(s|\eta)$ is treated as though it is a category traceline of a polytomous item (with $r_n + 1$ categories), and the Lord–Wingersky algorithm for polytomous item responses introduced in Section 2.5 is directly applied. Let $S_n^* = \sum_{j=1}^{n} r_j$ be the maximum summed score after item cluster n has been included in the recursions. To initialize, set the Step 1 summed score likelihood to the summed score likelihoods from the first cluster, i.e., $L_1\,(s|\eta) = L^1\,(s|\eta)$ for $s = 0,\ldots,S_1^*$. In step $n = 2,\ldots,N$, the summed score likelihoods from cluster n are added into the S_{n-1}^* available summed score likelihoods from the previous step:

$$L_n\,(s|\eta) = \sum_{s_*=0}^{S_{n-1}^*}\sum_{k=0}^{r_n} L_{n-1}\,(s_*|\eta)\,L^n\,(k|\eta)\,\mathbf{1}_s\,(s_*+k)\,, \tag{21}$$

where $\mathbf{1}_s\,(s_*+k)$ is still an indicator function that takes on a value of 1 if and only if s is equal to s_*+k, and 0 otherwise. Entirely analogous to Eq. (13), the summation in Eq. (21) is over the existing summed score likelihoods for scores $s_* = 0,\ldots,S_{n-1}^*$ and the $r_n + 1$ summed scores

TABLE 4.
Item parameters for the six dichotomous items with hypothetical bifactor structure.

Item	a^0	a^1	a^2	a^3	c
1	1.2	1.0			−1.0
2	1.2	1.0			−.6
3	1.0		.8		−.2
4	1.0		.8		.2
5	.8			1.2	.6
6	.8			1.2	1.0

from item cluster n, while preserving the restriction that the combination must lead to a summed score of s.

At the conclusion of step N, the likelihoods $L_N(s|\eta)$ are equal to the desired summed score likelihoods $L(s|\eta)$ for each s. Recall that $h(\eta)$ is the density of the primary dimension. Posterior summaries for summed score s can be readily computed using quadrature from $p(\eta|s) = L(s|\eta)\, h(\eta)/p(s)$, where the marginal probability $p(s) = \int L(s|\eta)\, h(\eta)\, \mathrm{d}\eta$ can be approximated with Q-point rectangular quadrature as $p(s) \approx \sum_{q=1}^{Q} L\left(s|X_q\right) W(X_q)$, with weights given by $W\left(X_q\right) = h(X_q)/\sum_{q=1}^{Q} h(X_q)$. Posterior mean and variance can be obtained with similar quadrature computations.

If there are more than one primary dimensions in the model or if any of the items are polytomous, the core structure of the algorithm remains the same. One would only have to replace the first-stage recursions in Eq. (19) by computations similar to those defined in Section 2.5, and use direct product quadrature rules for integrals over the vector-valued $\boldsymbol{\eta}$.

3.4. An Illustrative Example

Consider six hypothetical dichotomous items arranged in three doublets. There are four latent variables in this model, one primary dimension η on which all items load and three specific dimensions ξ_1, ξ_2, ξ_3. Table 4 shows the item parameters for these items, as well as the bifactor structure wherein items 1–2, 3–4, and 5–6 form into three doublets with non-zero loadings on the specific dimensions. The prior distributions of the latent variables are taken to be standard normal. Table 5 shows the ordinates of the item response functions as well as quadrature weights for the specific dimensions over a 5×5 grid defined by the direct product of equally spaced quadrature points at $-2, -1, 0, 1$, and 2. Due to space constraints, only values at a selected subset of the grid points are shown in Table 5. The weights for specific dimensions are normalized ordinates of standard normal densities as functions of ξ_1, ξ_2, and ξ_3, and repeated over the quadrature points for η. $W_1(\xi_1)$, $W_2(\xi_2)$, and $W_3(\xi_3)$ are the same in this example because the prior distributions of ξ_1, ξ_2, ξ_3 are all standard normal (but they need not always be standardized, see e.g., Cai et al., 2011).

Table 6 illustrates the first stage of the new recursive algorithm. In this case, summed score likelihoods are accumulated for each of the three item clusters. Within each item cluster, there are only two dichotomously scored items, so the summed scores range from 0 to 2. The summed score likelihoods are represented over separate grids formed by the direct product of the quadrature points for the primary dimension η crossed with ξ_1, ξ_2, and ξ_3, respectively. In Table 7, the specific dimensions are integrated out for each item cluster. This leaves the summed score likelihoods as functions of the primary dimension η alone.

Finally, the accumulated summed score likelihoods in each item cluster are used in the second stage of the recursive algorithm, as shown in Table 8. The within-cluster summed scores

TABLE 5.
Ordinates of item response functions and quadrature weights evaluated over the 5×5 direct product rectangular quadrature points for the six hypothetical items with bifactor structure.

η	−2	−2	−2	–	0	–	2	2	2
ξ_1	−2	−1	0	–	0	–	0	1	2
ξ_2	−2	−1	0	–	0	–	0	1	2
ξ_3	−2	−1	0	–	0	–	0	1	2
$W_1(\xi_1) =$	.054	.244	.403	–	.403	–	.403	.244	.054
$W_2(\xi_2) =$	.054	.244	.403	–	.403	–	.403	.244	.054
$W_3(\xi_3) =$	.054	.244	.403	–	.403	–	.403	.244	.054
Item 1: $T_1^1(1\|\eta, \xi_1) =$	.004	.012	.032	–	.269	–	.802	.917	.968
Item 2: $T_2^1(1\|\eta, \xi_1) =$	.007	.018	.047	–	.354	–	.858	.943	.978
Item 3: $T_1^2(1\|\eta, \xi_2) =$	.022	.047	.100	–	.450	–	.858	.931	.968
Item 4: $T_2^2(1\|\eta, \xi_2) =$	.032	.069	.142	–	.550	–	.900	.953	.978
Item 5: $T_1^3(1\|\eta, \xi_3) =$	.032	.100	.269	–	.646	–	.900	.968	.990
Item 6: $T_2^3(1\|\eta, \xi_3) =$	.047	.142	.354	–	.731	–	.931	.978	.993
Item 1: $T_1^1(0\|\eta, \xi_1) =$	.996	.988	.968	–	.731	–	.198	.083	.032
Item 2: $T_2^1(0\|\eta, \xi_1) =$	.993	.982	.953	–	.646	–	.142	.057	.022
Item 3: $T_1^2(0\|\eta, \xi_2) =$	.978	.953	.900	–	.550	–	.142	.069	.032
Item 4: $T_2^2(0\|\eta, \xi_2) =$	.968	.931	.858	–	.450	–	.100	.047	.022
Item 5: $T_1^3(0\|\eta, \xi_3) =$	.968	.900	.731	–	.354	–	.100	.032	.010
Item 6: $T_2^3(0\|\eta, \xi_3) =$	.953	.858	.646	–	.269	–	.069	.022	.007

are treated as though they are item scores for three polytomous items. At the end of the recursions the final summed score likelihoods for the primary dimension η are assembled and multiplied by the weights from the prior distribution of η, yielding posterior probabilities, expectations, and variances, as shown in Table 9. The entries under the heading posterior summaries form a summed score to IRT scaled score translation table (along with standard errors) for the primary dimension in an item bifactor model.

3.5. Some Additional Comparisons

Without the updated Lord–Wingersky algorithm, it may be tempting in practice to calibrate a test using a hierarchical item factor model (e.g., testlet model) to "handle" residual dependence, retain the general dimension slopes, and create a summed score to scaled score conversion table with the original unidimensional Lord–Wingersky algorithm. While this approach has a certain intuitive appeal, and the computation is simpler than the updated Lord–Wingersky algorithm, it is nevertheless going to lead to incorrect results. Failing to take into account the influence of residual dependence (as indicated by the presence of specific dimensions) in IRT scoring can still lead to an overstatement of the degree of reliability of the instrument. Recent work by Ip (2010a,b) and Stucky, Thissen, and Edelen (2013) also highlight the effects residual dependence has on scaled scores and standard errors.

Notably, the marginal reliability coefficient can become substantially overestimated. In the case of the illustrative example presented in Section 3.4, $\sigma^2(\eta)$ is equal to 1 because the prior $h(\eta)$ is standard normal. Applying Eq. (12) to results in Table 9, the marginal reliability of the scaled scores for the primary dimension η is equal to 0.47. On the other hand, if only the general dimension slopes in Table 4 are retained and standard Lord–Wingersky algorithm is applied to obtain a one-dimensional summed score conversion table (as shown in Table 10), the marginal

TABLE 6.
Accumulating summed score likelihoods within each item cluster.

		Quadrature grid for (η, ξ_1)								
Initialize cluster 1's summed score likelihoods by adding item 1										
Within-cluster	η	-2	-2	-2	–	0	–	2	2	2
Score likelihood	ξ_1	-2	-1	0	–	0	–	0	1	2
$L_1^1(0\|\eta,\xi_1) =$	$T_1^1(0\|\eta,\xi_1)$	.996	.988	0.968	–	.731	–	.198	.083	.032
$L_1^1(1\|\eta,\xi_1) =$	$T_1^1(1\|\eta,\xi_1)$	.004	.012	0.032	–	.269	–	.802	.917	.968
Add item 2 to cluster 1's summed score likelihoods										
$L_2^1(0\|\eta,\xi_1) =$	$L_1^1(0\|\eta,\xi_1)T_2^1(0\|\eta,\xi_1)$	.989	.970	.922	–	.472	–	.028	.005	.001
$L_2^1(1\|\eta,\xi_1) =$	$L_1^1(0\|\eta,\xi_1)T_2^1(1\|\eta,\xi_1)$ $+ L_1^1(1\|\eta,\xi_1)T_2^1(0\|\eta,\xi_1)$	.011	.030	.077	–	.433	–	.284	.131	.053
$L_2^1(2\|\eta,\xi_1) =$	$L_1^1(1\|\eta,\xi_1)T_2^1(1\|\eta,\xi_1)$	.000	.000	.002	–	.095	–	.688	.864	.947
		Quadrature grid for (η, ξ_2)								
Initialize cluster 2's summed score likelihoods by adding item 3										
Within-cluster	η	-2	-2	-2	–	0	–	2	2	2
Score likelihood	ξ_2	-2	-1	0	–	0	–	0	1	2
$L_1^2(0\|\eta,\xi_2) =$	$T_1^2(0\|\eta,\xi_1)$	.978	.953	.90	–	.550	–	.142	.069	.032
$L_1^2(1\|\eta,\xi_2) =$	$T_1^2(1\|\eta,\xi_1)$	.022	.047	.10	–	.450	–	.858	.931	.968
Add item 4 to cluster 2's summed score likelihoods										
$L_2^2(0\|\eta,\xi_2) =$	$L_1^2(0\|\eta,\xi_2)T_2^2(0\|\eta,\xi_2)$	.947	.887	.773	–	.248	–	.014	.003	.001
$L_2^2(1\|\eta,\xi_2) =$	$L_1^2(0\|\eta,\xi_2)T_2^2(1\|\eta,\xi_2)$ $+ L_1^2(1\|\eta,\xi_2)T_2^2(0\|\eta,\xi_2)$	.053	.110	.213	–	.505	–	.213	.110	.053
$L_2^2(2\|\eta,\xi_2) =$	$L_1^2(1\|\eta,\xi_2)T_2^2(1\|\eta,\xi_2)$	.001	.003	.014	–	.248	–	.773	.887	.947
		Quadrature grid for (η, ξ_3)								
Initialize cluster 3's summed score likelihoods by adding item 5										
Within-cluster	η	-2	-2	-2	–	0	–	2	2	2
Score likelihood	ξ_3	-2	-1	0	–	0	–	0	1	2
$L_1^3(0\|\eta,\xi_3) =$	$T_1^3(0\|\eta,\xi_3)$	.968	.900	.731	–	.354	–	.100	.032	.010
$L_1^3(1\|\eta,\xi_3) =$	$T_1^3(1\|\eta,\xi_3)$	.032	.100	.269	–	.646	–	.900	.968	.990
Add item 6 to cluster 3's summed score likelihoods										
$L_2^3(0\|\eta,\xi_3) =$	$L_1^3(0\|\eta,\xi_3)T_2^3(0\|\eta,\xi_3)$	.922	.773	.472	–	.095	–	.007	.001	.000
$L_2^3(1\|\eta,\xi_3) =$	$L_1^3(0\|\eta,\xi_3)T_2^3(1\|\eta,\xi_3)$ $+ L_1^3(1\|\eta,\xi_3)T_2^3(0\|\eta,\xi_3)$	.077	.213	.433	–	.433	–	.155	.053	.017
$L_2^3(2\|\eta,\xi_3) =$	$L_1^3(1\|\eta,\xi_3)T_2^3(1\|\eta,\xi_3)$	.002	.014	.195	–	.472	–	.838	.947	.983

reliability of the scaled scores for summed scores becomes 0.56, an almost 20 % upward bias relative to the reliability estimate from the more appropriate scoring method.

Furthermore, the estimates of scaled scores are also impacted. A comparison between Tables 9 and 10 shows that the posterior means become more extreme in general when the specific dimension slopes are ignored and the unidimensional scoring algorithm used. This is natural since the item intercepts and slopes are unstandardized parameters. When the (typically positive) specific dimension slopes are ignored and the intercepts remain untouched, the implied standardized threshold parameters becomes more extreme, leading to posteriors that are positioned more toward the extreme ends of the latent trait scale.

TABLE 7.
Integrating the specific dimensions out of the summed score likelihoods.

	Quadrature grid for (η, ξ_1)								
η	−2	−2	−2	–	0	–	2	2	2
ξ_1	−2	−1	0	–	0	–	0	1	2
Multiply cluster 1's summed score likelihoods by $W_1(\xi_1)$									
$L^1(0\|\eta,\xi_1)W_1(\xi_1) = L_2^1(0\|\eta,\xi_1)W_1(\xi_1)$	.054	.237	.371	–	.190	–	.011	.001	.000
$L^1(1\|\eta,\xi_1)W_1(\xi_1) = L_2^1(1\|\eta,\xi_1)W_1(\xi_1)$	.001	.007	.031	–	.174	–	.114	.032	.003
$L^1(2\|\eta,\xi_1)W_1(\xi_1) = L_2^1(2\|\eta,\xi_1)W_1(\xi_1)$	.000	.000	.001	–	.038	–	.277	.211	.052

	η				
	−2	−1	0	1	2
Summing over ξ_1, leaving cluster 1's summed score likelihoods as functions of η only					
$L^1(0\|\eta) = \sum_{\xi_1} L^1(0\|\eta,\xi_1)W_1(\xi_1)$	.891	.728	.469	.212	.062
$L^1(1\|\eta) = \sum_{\xi_1} L^1(1\|\eta,\xi_1)W_1(\xi_1)$	.103	.235	.382	.411	.288
$L^1(2\|\eta) = \sum_{\xi_1} L^1(2\|\eta,\xi_1)W_1(\xi_1)$	.006	.037	.148	.377	.649

	Quadrature grid for (η, ξ_2)								
η	−2	−2	−2	–	0	–	2	2	2
ξ_2	−2	−1	0	–	0	–	0	1	2
Multiply cluster 2's summed score likelihoods by $W_2(\xi_2)$									
$L^2(0\|\eta,\xi_2)W_2(\xi_2) = L_2^2(0\|\eta,\xi_2)W_2(\xi_2)$	.052	.217	.311	–	.100	–	.006	.001	.000
$L^2(1\|\eta,\xi_2)W_2(\xi_2) = L_2^2(1\|\eta,\xi_2)W_2(\xi_2)$	.003	.027	.086	–	.203	–	.086	.027	.003
$L^2(2\|\eta,\xi_2)W_2(\xi_2) = L_2^2(2\|\eta,\xi_2)W_2(\xi_2)$	.000	.001	.006	–	.100	–	.311	.217	.052

	η				
	−2	−1	0	1	2
Summing over ξ_2, leaving cluster 2's summed score likelihoods as functions of η only					
$L^2(0\|\eta) = \sum_{\xi_2} L^2(0\|\eta,\xi_2)W_2(\xi_2)$	.742	.519	.277	.106	.028
$L^2(1\|\eta) = \sum_{\xi_2} L^2(1\|\eta,\xi_2)W_2(\xi_2)$	.230	.375	.446	.375	.230
$L^2(2\|\eta) = \sum_{\xi_2} L^2(2\|\eta,\xi_2)W_2(\xi_2)$	.028	.106	.277	.519	.742

	Quadrature grid for (η, ξ_3)								
η	−2	−2	−2	–	0	–	2	2	2
ξ_3	−2	−1	0	–	0	–	0	1	2
Multiply cluster 3's summed score likelihoods by $W_3(\xi_3)$									
$L^3(0\|\eta,\xi_3)W_3(\xi_3) = L_2^3(0\|\eta,\xi_3)W_3(\xi_3)$	.050	.189	.190	–	.038	–	.003	.000	.000
$L^3(1\|\eta,\xi_3)W_3(\xi_3) = L_2^3(1\|\eta,\xi_3)W_3(\xi_3)$	.004	.052	.174	–	.174	–	.062	.013	.001
$L^3(2\|\eta,\xi_3)W_3(\xi_3) = L_2^3(2\|\eta,\xi_3)W_3(\xi_3)$	.000	.003	.038	–	.190	–	.337	.231	.054

TABLE 7.
continued

	η				
	−2	−1	0	1	2
Summing over ξ_3, leaving cluster 1's summed score likelihoods as functions of η only					
$L^3(0\|\eta) = \sum_{\xi_3} L^3(0\|\eta, \xi_3) W_3(\xi_3)$	.469	.302	.166	.077	.029
$L^3(1\|\eta) = \sum_{\xi_3} L^3(1\|\eta, \xi_3) W_3(\xi_3)$	.364	.396	.364	.285	.192
$L^3(2\|\eta) = \sum_{\xi_3} L^3(2\|\eta, \xi_3) W_3(\xi_3)$	.166	.302	.469	.638	.779

TABLE 8.
Forming summed score likelihoods for the primary dimension.

Summed score likelihoods	η	−2	−1	0	1	2
Step 1: initialize summed score likelihoods by adding item cluster 1						
$L_1(0\|\eta) =$	$L^1(0\|\eta)$	.891	.728	.469	.212	.062
$L_1(1\|\eta) =$	$L^1(1\|\eta)$	.103	.235	.382	.411	.288
$L_1(2\|\eta) =$	$L^1(2\|\eta)$	.006	.037	.148	.377	.649
Step 2: add item cluster 2 to existing summed score likelihoods						
$L_2(0\|\eta) =$	$L^1(0\|\eta)\, L^2(0\|\eta)$	.661	.378	.130	.022	.002
$L_2(1\|\eta) =$	$L^1(0\|\eta)\, L^2(1\|\eta) + L^1(1\|\eta)\, L^2(0\|\eta)$	.281	.395	.315	.123	.022
$L_2(2\|\eta) =$	$L^1(0\|\eta)\, L^2(2\|\eta) + L^1(1\|\eta)\, L^2(1\|\eta) + L^1(2\|\eta)\, L^2(0\|\eta)$	.053	.184	.342	.304	.131
$L_2(3\|\eta) =$	$L^1(1\|\eta)\, L^2(2\|\eta) + L^1(2\|\eta)\, L^2(1\|\eta)$	.004	.039	.172	.355	.363
$L_2(4\|\eta) =$	$L^1(2\|\eta)\, L^2(2\|\eta)$	.000	.004	.041	.196	.482
Step 3: add item cluster 3 to existing summed score likelihoods						
$L_3(0\|\eta) =$	$L^2(0\|\eta)\, L^3(0\|\eta)$	.310	.114	.022	.002	.000
$L_3(1\|\eta) =$	$L^2(0\|\eta)\, L^3(1\|\eta) + L^2(1\|\eta)\, L^3(0\|\eta)$	.373	.269	.100	.016	.001
$L_3(2\|\eta) =$	$L^2(0\|\eta)\, L^3(2\|\eta) + L^2(1\|\eta)\, L^3(1\|\eta) + L^2(2\|\eta)\, L^3(0\|\eta)$	.237	.326	.233	.073	.010
$L_3(3\|\eta) =$	$L^2(1\|\eta)\, L^3(2\|\eta) + L^2(2\|\eta)\, L^3(1\|\eta) + L^2(3\|\eta)\, L^3(0\|\eta)$	.068	.204	.301	.192	.053
$L_3(4\|\eta) =$	$L^2(2\|\eta)\, L^3(2\|\eta) + L^2(3\|\eta)\, L^3(1\|\eta) + L^2(4\|\eta)\, L^3(0\|\eta)$	.010	.072	.230	.310	.186
$L_3(5\|\eta) =$	$L^2(3\|\eta)\, L^3(2\|\eta) + L^2(4\|\eta)\, L^3(1\|\eta)$	.001	.013	.096	.282	.375
$L_3(6\|\eta) =$	$L^2(4\|\eta)\, L^3(2\|\eta)$	.000	.001	.019	.125	.375

4. Additional Applications

Besides summed score based IRT scoring tables, the updated Lord–Wingersky algorithm can be applied creatively to solve a test linking problem (see Thissen et al., 2011), to create score combination tables for mixed-format tests, and to construct model fit test statistics. Discussed in this section are only selections of the new possibilities opened up by the updated algorithm.

4.1. Calibrated Projection Linking

Thissen et al. (2011) described a novel test linking method called calibrated projection that fuses simultaneous calibration with projection linking. The main advantage of calibrated projection is its ability to link two closely related (though not conceptually identical) scales in a single

TABLE 9.
Characterizing the summed score likelihoods and posteriors for the primary dimension.

Quadrature	η							
Weights at	−2	−1	0	1	2			
$W(\eta) =$	.054	.244	.403	.244	.054			
Summed score	η							
Likelihoods $L(s\|\eta)$ at	−2	−1	0	1	2			
$L(0\|\eta) =$	.310	.114	.022	.002	.000			
$L(1\|\eta) =$	.373	.269	.100	.016	.001			
$L(2\|\eta) =$	.237	.326	.233	.073	.010			
$L(3\|\eta) =$	.068	.204	.301	.192	.053			
$L(4\|\eta) =$	.010	.072	.230	.310	.186			
$L(5\|\eta) =$	.001	.013	.096	.282	.375			
$L(6\|\eta) =$	.000	.001	.019	.125	.375			
Unnormalized summed	η					Posterior summaries		
Score posteriors $p(\eta\|s)$ at	−2	−1	0	1	2	$p(s)$	$E(\eta\|s)$	$V(\eta\|s)$
$p(\eta\|0) \propto L(0\|\eta)W(\eta) =$	.017	.028	.009	.000	.000	.05	−1.14	.49
$p(\eta\|1) \propto L(1\|\eta)W(\eta) =$	.020	.066	.040	.004	.000	.13	−.79	.54
$p(\eta\|2) \propto L(2\|\eta)W(\eta) =$	.013	.080	.094	.018	.001	.20	−.42	.56
$p(\eta\|3) \propto L(3\|\eta)W(\eta) =$	.004	.050	.121	.047	.003	.22	−.02	.55
$p(\eta\|4) \propto L(4\|\eta)W(\eta) =$	.001	.018	.093	.076	.010	.20	.39	.54
$p(\eta\|5) \propto L(5\|\eta)W(\eta) =$	.000	.003	.039	.069	.020	.13	.81	.52
$p(\eta\|6) \propto L(6\|\eta)W(\eta) =$	.000	.000	.008	.030	.020	.06	1.21	.46

TABLE 10.
Summed score to scaled score conversions based on primary dimension slopes only.

Summed	Posterior summaries		
Scores	$p(s)$	$E(\eta\|s)$	$V(\eta\|s)$
$s = 0$	.05	−1.29	.40
$s = 1$	.13	−.90	.46
$s = 2$	.20	−.47	.46
$s = 3$	.22	−.03	.44
$s = 4$	.20	.42	.44
$s = 5$	.14	.89	.43
$s = 6$	.06	1.33	.37

step that is entirely based on multidimensional IRT calibration. Thissen et al. (2011) illustrated the application of calibrated projection in health outcomes research, wherein a legacy instrument (PedsQL™ Asthma Symptoms Module) was projection linked onto the scale of the new pediatric asthma impact scale (PAIS). PAIS was built with IRT methods, whereas PedsQL™ was built with classical test theory methods, thus requiring the use of summed scoring. Producing a scoring cross-walk would enable the clinicians and researchers who already use PedQL™ to report scaled scores comparable to PAIS.

As illustrated by Thissen et al.'s (2011) Tables 2 and 3, both instruments use five-point ordered response scales suitable for the graded response model and each may be considered approximately unidimensional. PedsQL™ Asthma Symptoms Module contains 11 items and PAIS has 17. A

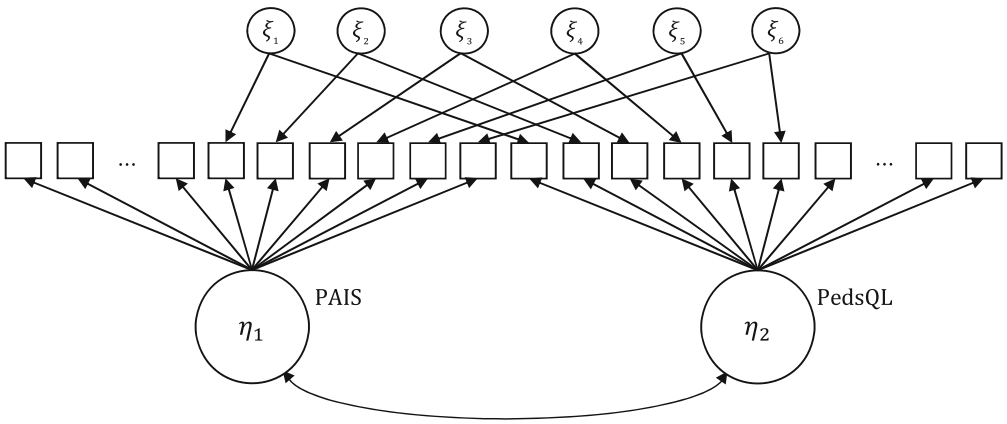

FIGURE 2.
Path diagram of a two-tier model for calibrated projection linking. The two primary dimensions are correlated at .96 and there are six item doublets.

multitude of additional differences between the two instruments implies that the more stringent requirements of concurrent calibration (e.g., equal construct) are probably not satisfied. Hence the weaker prediction/projection methods must be employed.

At the core of calibrated projection linking is a multidimensional IRT model that has at least two correlated primary dimensions (η_1 and η_2), each measured by the respective instrument (PAIS and PedsQL™) with an independent cluster factor pattern. The correlation between η_1 and η_2 is estimated simultaneously with the item parameters. The multidimensional IRT model then produces scores (projected through the correlation) on the scale of one instrument (PAIS in this case) using only the responses to items from the other instrument (PedsQL™ Asthma Symptoms Module). This model, when depicted in a graph, resembles the bottom half of the path diagram shown in Figure 2.

However, when the two instruments were considered together, strong local dependence emerged among six pairs of items. As it turns out, these six pairs of items have stem wording that are virtually identical. For example, item 13 of PAIS reads "I had asthma attacks," and item 20 of PedsQL™ Asthma Symptoms Module reads "I have asthma attacks." The six items in fact represent some of the best symptoms that are indicative of asthma's impact. Consequently, Thissen et al. (2011) suggested including six orthogonal latent variables to account for the effects of local dependence. This model is depicted in Figure 2. It is formally a two-tier model with $M = 2$ primary dimensions and $N = 6$ specific dimensions. The two primary dimensions are assumed to be bivariate normal, standardized in each dimension, with an unknown correlation coefficient. Thissen et al. (2011) obtained a linking sample and estimated the correlation coefficient ($r = 0.96$) as well as the item parameters for both instruments.

Retaining the item parameters for PedsQL™ reported in Thissen et al. (2011), it is straightforward to apply the updated Lord–Wingersky algorithm. Table 11 shows the item parameters for the 11 PedsQL™ items. The slopes on the first general dimension η_1, representing PAIS, are all equal to zero here, indicating the absence of items that cross-load on both dimensions. The PAIS item slopes do not enter into the projection linking computations because only items from PedsQL™ are considered (along with the 0.96 prior correlation). The non-zero slopes for the six specific dimensions (ξ_1–ξ_6) are what remain of the item doublet slopes after removing their counterparts among the PAIS items.

For each summed score ($s = 0, \ldots, 44$) on PedsQL™, the recursive algorithm produces a bivariate posterior for η_1 and η_2. Figure 3 shows the bivariate normal approximations to three

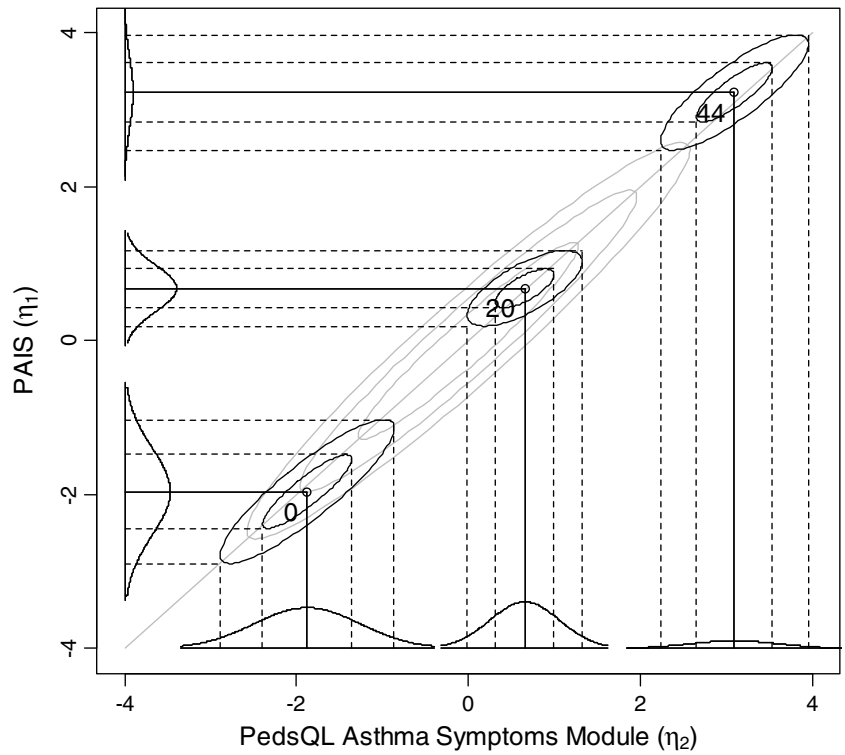

FIGURE 3.
Bivariate contour plots showing three selected summed score posteriors for PedsQL™ Asthma Symptoms Module as well as the projected posteriors on the PAIS scale.

selected posteriors, for summed scores 0, 20, and 44, overlaid on the gray contours representing the bivariate normal prior with an estimated correlation of 0.96. The x-axis of Figure 3 represents the PedsQL™ latent variable (η_2), whereas the y-axis represents PAIS (η_1), consistent with the notation in Figure 2. The marginal posteriors are also plotted, indicating that entire summed score posteriors are projected through the bivariate relation between η_1 and η_2. The marginal posteriors on the y-axis are of key interest. Their relative sizes indicate the model-implied summed score proportions. Their means and variances become scores and error variances on the scale of PAIS for each PedsQL™ summed score, corrected for local dependence.

4.2. Score Combination

Modern educational assessments are often made up of items of varying types. For instance, a test may consist of traditional MC items that are dichotomously scored, for which the classical three-parameter IRT model may be useful, as well as items that require judge-rated CRs or performance tasks that are subsequently analyzed using the graded response model (Samejima, 1969) or the generalized partial credit model (Muraki, 1992). When the MC items and the CR items measure the same latent construct and the test is approximately unidimensional, reporting a single combined score is a sensible approach. Rosa et al. (2001) proposed a score combination method that is based on the pattern of summed scores from the MC and CR sections. This is a convenient and practical approximation to the optimal (but more involved) scoring with the full response pattern.

TABLE 11.
Item parameters for the 11 PedsQL™ items as input into the Lord–Wingersky 2.0 algorithm.

Item	Slopes								Intercepts			
	η_1	η_2	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	1	2	3	4
1	0	2.31	0	0	0	0	0	0	.77	−.56	−3.19	−5.50
2	0	3.90	0	2.37	0	0	0	0	1.50	−1.05	−5.83	−8.24
3	0	4.09	3.85	0	0	0	0	0	−2.04	−4.89	−9.10	−12.15
4	0	1.70	0	0	0	0	0	0	−.48	−1.20	−2.84	−3.68
5	0	2.25	0	0	0	0	0	0	2.05	.69	−2.14	−3.82
6	0	2.63	0	0	0	0	0	2.52	4.44	2.17	−1.70	−4.08
7	0	3.42	0	0	2.04	0	0	0	1.79	−.65	−4.59	−7.02
8	0	1.07	0	0	0	0	0	0	1.64	.55	−1.29	−2.29
9	0	3.11	0	0	0	0	1.66	0	−.17	−1.88	−4.11	−5.82
10	0	3.36	0	0	0	4.06	0	0	−1.91	−4.02	−7.34	−9.21
11	0	2.19	0	0	0	0	0	0	.14	−1.18	−3.44	−5.02

Specifically, let the summed score likelihoods for the MC section be $L_{\mathrm{MC}}(s|\theta)$, and $s = 0, \ldots, S_{\mathrm{MC}}$, where S_{MC} is the maximum summed score for the MC section. Similarly, let $L_{\mathrm{CR}}(s|\theta), s = 0, \ldots, S_{\mathrm{CR}}$ denote the summed score likelihoods for the CR section. Rosa et al. (2001) states that following summed score pattern posterior provides a basis for combining MC section score s_1 with CR section score s_2:

$$p(\theta|s_1, s_2) = \frac{L_{\mathrm{MC}}(s_1|\theta)\, L_{\mathrm{CR}}(s_2|\theta)\, g(\theta)}{\int L_{\mathrm{MC}}(s_1|\theta)\, L_{\mathrm{CR}}(s_2|\theta)\, g(\theta)\, \mathrm{d}\theta}. \tag{22}$$

To compute the posterior, Rosa et al. (2001) noted that one would have to apply the standard Lord–Wingersky algorithm to the two sections separately and then explicitly use Eq. (22) to construct a two-way look-up table for each of the summed score patterns.

If one regards the MC section as a testlet, and the CR section as another one, one may choose to rewrite Eq. (22) as:

$$p(\theta|s_1, s_2) = \frac{\int L_{\mathrm{MC}}(s_1|\theta)\, g_1(\xi_1)\, \mathrm{d}\xi_1 \int L_{\mathrm{CR}}(s_2|\theta)\, g_2(\xi_2)\, \mathrm{d}\xi_2 g(\theta)}{\int\int L_{\mathrm{MC}}(s_1|\theta)\, g_1(\xi_1)\, \mathrm{d}\xi_1 \int L_{\mathrm{CR}}(s_2|\theta)\, g_2(\xi_2)\, \mathrm{d}\xi_2 g(\theta)\, \mathrm{d}\theta}. \tag{23}$$

Note that the key condition for $p(\theta|s_1, s_2)$ in Eq. (22) to be the same as Eq. (23) is: $L_{\mathrm{MC}}(s_1|\theta, \xi_1) = L_{\mathrm{MC}}(s_1|\theta)$ and $L_{\mathrm{CR}}(s_2|\theta, \xi_2) = L_{\mathrm{CR}}(s_2|\theta)$. In other word, the two are the same when items in both MC and CR sections do not depend on the specific dimensions ξ_1 and ξ_2; or, alternatively, when the item slopes on ξ_1 and ξ_2 are all equal to zero. The equivalence suggests that one does not need a specialized algorithm for implementing Rosa et al.'s (2001) scoring combination method. One would simply have to set up a special bifactor model wherein all specific dimension slopes are constrained to zero and apply Version 2.0 of the Lord–Wingersky algorithm outlined in Section 3 to this bifactor model. Although the specific dimension slopes may be zero, the presence of the testlet structure enables the first stage of the updated Lord–Wingersky algorithm to accumulate the within-section summed score likelihoods separately. Instead of collapsing the section-specific summed scores as per Eq. (21), the pattern of summed scores is used to compute a posterior for the primary dimension directly.

As a concrete example, consider the Wisconsin 3rd grade reading assessment items discussed in Rosa et al. (2001). There are altogether 20 items, 16 in the MC section (scored 0-1), and four in

TABLE 12.
Item parameters for the 20 Wisconsin 3rd grade reading items as input into the Lord-Wingersky 2.0 algorithm.

Item	Slopes			Intercept	Guessing	
	θ	ξ_1	ξ_2			
Multiple-choice items (3PL Model)						
1	1.02	0	0	.72	.20	
2	2.16	0	0	2.99	.31	
3	2.29	0	0	2.72	.22	
4	1.47	0	0	1.37	.23	
5	2.29	0	0	.92	.23	
6	3.61	0	0	1.83	.19	
7	2.05	0	0	1.12	.23	
8	2.60	0	0	3.36	.28	
9	1.47	0	0	1.36	.20	
10	2.76	0	0	1.68	.18	
11	1.88	0	0	1.84	.22	
12	2.27	0	0	.84	.28	
13	1.46	0	0	1.11	.20	
14	3.90	0	0	1.81	.25	
15	1.56	0	0	.14	.26	
16	1.62	0	0	2.02	.21	
Item	Slopes			Intercepts		
	θ	ξ_1	ξ_2	1	2	3
Rated constructed response items (graded model)						
1	.87	0	0	4.29	2.48	-1.01
2	.93	0	0	4.15	1.33	-1.06
3	1.31	0	0	4.47	2.31	.69
4	.73	0	0	4.05	1.27	-1.63

the CR section (each has four score points). Using the item parameters reported by Thissen and Wainer (2001), one may set up a bifactor model with two empty specific dimensions (as shown in Table 12). Application of the updated Lord–Wingersky algorithm to the model in Table 12 leads to a two-way table (Table 13) that (almost) reproduces Table 7.2 (p. 259) in Rosa et al. (2001) with any difference attributable to limited number of significant digits in the reported item parameters and numerical quadrature error. For instance, the summed score combination EAP score for a student who scored 14 items correct in the MC section and received a CR score of 7 is equal to .01. A by-product of the computation of the summed score combination EAPs is the summed score pattern probabilities. With the variously shaded areas in Table 13, the probabilities associated with each of the two-way combinations are utilized to indicate the highest density regions (HDR) of the summed score combination posterior. Detailed procedures for constructing these HDRs are described in Thissen and Wainer (2001, p. 260), which involves finding those summed score combinations that belong to the top x % of the cumulative probability distribution of the two-way combinations. The unshaded cells represent collectively the 99 % HDR, and the lightly shaded ones are the 99.9 % HDR. As Thissen and Wainer (2001) noted, the shaded cells represent those summed score combinations that occur very rarely, e.g., a perfect CR section score of 12 crossed with any MC section score of less than 11. Rather than proceeding with scoring the "aberrant" pattern, the additional information contained in the table may provide useful diagnostics for the testing program.

TABLE 13.
Summed score combination table computed by the updated recursive algorithm for the Wisconsin reading items.

Summed Score for MC Items	Summed Rated Score for CR Items												
	0	1	2	3	4	5	6	7	8	9	10	11	12
0	-3.28	-3.05	-2.84	-2.66	-2.50	-2.35	-2.22	-2.11	-2.01	-1.92	-1.85	-1.79	-1.73
1	-3.23	-2.98	-2.77	-2.58	-2.42	-2.27	-2.13	-2.01	-1.91	-1.82	-1.75	-1.68	-1.62
2	-3.17	-2.91	-2.69	-2.50	-2.32	-2.17	-2.03	-1.91	-1.80	-1.71	-1.63	-1.57	-1.51
3	-3.10	-2.83	-2.59	-2.39	-2.22	-2.06	-1.92	-1.79	-1.68	-1.59	-1.51	-1.45	-1.38
4	-3.01	-2.72	-2.48	-2.27	-2.09	-1.93	-1.79	-1.66	-1.56	-1.46	-1.38	-1.32	-1.25
5	-2.90	-2.59	-2.34	-2.12	-1.94	-1.78	-1.64	-1.52	-1.42	-1.33	-1.25	-1.18	-1.12
6	-2.75	-2.43	-2.16	-1.95	-1.77	-1.62	-1.49	-1.37	-1.27	-1.19	-1.11	-1.05	-.99
7	-2.55	-2.21	-1.95	-1.75	-1.59	-1.45	-1.33	-1.22	-1.13	-1.05	-.98	-.91	-.86
8	-2.29	-1.95	-1.71	-1.53	-1.39	-1.27	-1.16	-1.07	-.98	-.90	-.83	-.77	-.72
9	-1.94	-1.64	-1.44	-1.30	-1.18	-1.08	-.99	-.91	-.83	-.76	-.69	-.63	-.57
10	-1.54	-1.32	-1.18	-1.07	-.98	-.90	-.82	-.75	-.67	-.60	-.53	-.47	-.41
11	-1.15	-1.02	-.93	-.85	-.78	-.72	-.65	-.58	-.51	-.44	-.37	-.30	-.23
12	-.83	-.76	-.70	-.65	-.59	-.53	-.47	-.40	-.33	-.25	-.18	-.09	-.01
13	-.57	-.53	-.49	-.44	-.39	-.34	-.28	-.21	-.13	-.05	.05	.16	.27
14	-.33	-.30	-.27	-.23	-.18	-.13	-.07	.01	.10	.20	.33	.47	.63
15	-.10	-.08	-.04	.00	.05	.11	.18	.27	.38	.51	.67	.87	1.11
16	.15	.18	.21	.26	.32	.39	.48	.59	.72	.89	1.11	1.37	1.70

While the foregoing may be deemed a convenient trick for tests that are unidimensional, it does offer a degree of generality that Rosa et al.'s (2001) original method did not possess. That is, when the MC or CR sections demonstrate departures from unidimensionality, e.g., when there is testing mode effect for the CR items, and the specific slopes may not be exactly zero, the new algorithm will properly adjust the combined scaled score for residual dependence, requiring no new specialized implementation.

4.3. Model Fit Evaluation

As soon as summed score probabilities can be evaluated for unidimensional IRT models, researchers have explored their use in model fit diagnosis. Orlando and Thissen's (2000) summed score likelihood based item fit statistic is one prominent example. Described here is a generalized version of the summed score fit statistic implemented in flexMIRT® (Cai, 2013). Consider polytomous items $i = 1, \ldots, I$ with K_i categories. Recall that the maximum summed score is $S = \sum_{i=1}^{I}(K_i - 1)$. One may compute the "rest score" likelihoods, i.e., the summed score likelihoods based on all items except i. Let $L_{(i)}(s|\theta)$, $s = 0, \ldots, S - (K_i - 1)$, denote the rest score likelihoods for item i. For this item, the probability for category k in rest score group s is

$$p_{ik}(s) = \int L_{(i)}(s|\theta)\, T_i(k|\theta)\, g(\theta)\, \mathrm{d}\theta. \tag{24}$$

The probability for rest score group s is

$$p_{(i)}(s) = \int L_{(i)}(s|\theta)\, g(\theta)\, \mathrm{d}\theta. \tag{25}$$

Therefore, the model-implied probability of endorsing category k if the rest score is s can be computed as $E_{ik}(s) = p_{ik}(s)/p_{(i)}(s)$. The observed probability of endorsing category k if the rest score is s can be found by tabulating the calibration data. Let it be denoted as $O_{ik}(s)$. A Pearson-type statistic may be constructed as follows:

$$S - X_i^2 = \text{Sample size} \times \sum_{s=0}^{S-(K_i-1)} O_{(i)}(s) \sum_{k=0}^{K_i-1} \frac{(O_{ik}(s) - E_{ik}(s))^2}{E_{ik}(s)(1 - E_{ik}(s))}, \tag{26}$$

where $O_{(i)}(s)$ is the observed counterpart to $p_{(i)}(s)$. Orlando and Thissen (2000) presented simulation evidence that the large sample distribution of $S - X_i^2$ (at least in the dichotomous case) can be approximated by a central χ^2 distribution with $S - (K_i - 1) - q_i$ degrees-of-freedom, where q_i is the number of freely estimated item parameters for item i.

With the updated Lord–Wingersky algorithm, it is straightforward to generalize $S - X^2$ to hierarchical item factor models. Some additional book-keeping is necessary, however, to fully utilize dimension reduction. Consider item i in cluster/testlet n. Let $L_{(n)}(s|\boldsymbol{\eta})$ denote the summed score likelihoods in terms of the primary dimensions $\boldsymbol{\eta}$, accumulated over all item clusters other than cluster n. $L_{(n)}(s|\boldsymbol{\eta})$ is straightforward to compute by ignoring cluster n after stage 1 of the recursions is completed. Recall that r_n is the maximum within-cluster score for cluster n. Thus $L_{(n)}(s|\boldsymbol{\eta})$ is defined for $s = 0, \ldots, S - r_n$. Within cluster n, let the summed score likelihoods without item i be $L_{(n)}^{(i)}(s|\boldsymbol{\eta}, \xi_n)$. Note that the dependence on specific dimension is not yet integrated out of the likelihood, and $L_{(n)}^{(i)}(s|\boldsymbol{\eta}, \xi_n)$ is defined for $s = 0, \ldots r_n - (K_i - 1)$.

The posterior probability for category k in rest score group s is

$$\begin{aligned} p_{ik}(s) = &\int \sum_{s_1=0}^{S-r_n} L_{(n)}(s_1|\boldsymbol{\eta}) \\ &\int \sum_{s_2=0}^{r_n-(K_i-1)} L_{(n)}^{(i)}(s_2|\boldsymbol{\eta}, \xi_n)\, \mathbf{1}_s(s_1 + s_2)\, T_i(k|\boldsymbol{\eta}, \xi_n)\, g_n(\xi_n)\, \mathrm{d}\xi_n h(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}, \end{aligned} \tag{27}$$

where $\mathbf{1}_s(s_1 + s_2)$ is an indicator function that is equal to 1 if and only if $s = s_1 + s_2$, and 0 otherwise. The inner summation is needed because it combines likelihoods from cluster n while enforcing the constraint that the rest score must be s, before the dependence on specific dimension n is integrated out. By analogy, the posterior probability for rest score group s is

$$p_{(i)}(s) = \int \sum_{s_1=0}^{S-r_n} L_{(n)}(s_1|\boldsymbol{\eta}) \int \sum_{s_2=0}^{r_n-(K_i-1)} L_{(n)}^{(i)}(s_2|\boldsymbol{\eta}, \xi_n)\, \mathbf{1}_s(s_1 + s_2)\, g_n(\xi_n)\, \mathrm{d}\xi_n h(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}. \tag{28}$$

Once the posterior probabilities are computed, they can be inserted into Eq. (26) to evaluate a χ^2 test statistic for item i. Li and Rupp (2011) examined a version of this index by simulation but did not discuss the recursive algorithm that is needed to compute $S - X^2$ for hierarchical item factor models in full generality.

Finally, the model-implied summed score probabilities themselves, when compared against the observed probabilities, may be useful for diagnosing the ubiquitous latent variable normality assumption for the primary dimension in a testlet or bifactor model. While the idea itself is not new (see Ferrando & Lorenzo-seva, 2001; Hambleton & Traub, 1973; Lord, 1953; Ross, 1966; Sinharay, Johnson, & Stern, 2006), its use in hierarchical item factor models requires the new Lord–Wingersky algorithm (Li & Cai, 2012).

5. Discussion

Hierarchical item factor models can relax some of the restrictive assumptions of unidimensional IRT models. They have been suggested as useful tools for educational and psychological measurement research and practice in that they may better reflect the structure of measurement instruments (Reise, 2012). They respect the fact that many constructs have multi-faceted manifestations and yet it remains desirable to report on a single composite scale. This is especially true with measurement in mental health (e.g., depression symptoms), but also true in educational assessment (e.g., language testing). The multitude of group/specific factors in hierarchical item factor models can appropriately accommodate the inherent multidimensionality without altering the interpretability of the general dimension.

The mathematical complexity of hierarchical item factor models, however, makes their routine use unrealistic. Importantly, scoring tests with bifactor/testlet/two-tier models can be computational involving and specialized software programs are required if response pattern scores are needed. Utilizing dimension reduction, an updated Lord–Wingerksy algorithm is presented in this paper. This algorithm is computationally efficient even under a large number of latent factors.

With the updated Lord–Wingersky algorithm, one may adopt a hierarchical item factor model in the test design and item calibration stage and produce summed score conversions that are as convenient to use in practical settings as the original Lord–Wingersky method. The conversion tables are properly adjusted for the effects of residual dependence. To the end-user, the conversion tables eliminated the scoring complexities associated with the adoption of a multidimensional measurement model. Once the table is assembled, no specialized software is necessary for the end-user to reap the benefits of hierarchical multidimensional IRT modeling, thereby eliminating one of the key barriers to more wide-spread applications of hierarchical item factor models. In addition, the new algorithm serves as the basis of new test linking methods (calibrated projection), encompass traditional score combination approaches, and lead to new model fit diagnostic statistics. The new algorithm is fully implemented in IRTPRO (Cai et al., 2011) and flexMIRT® (Cai, 2013).

Acknowledgments

Part of this research is supported by the Institute of Education Sciences (R305B080016 and R305D100039) and the National Institute on Drug Abuse (R01DA026943 and R01DA030466). The views expressed here belong to the author and do not reflect the views or policies of the funding agencies. The author is grateful to Dr. David Thissen and members of the UCLA psychometric lab (in particular Carl Falk, Jane Li, and Ji Seung Yang) for comments on an earlier draft.

References

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L. (2013). *flexMIRT Version 2.0: Flexible multilevel item analysis and test scoring (Computer software)*. Chapel Hill, NC: Vector Psychometric Group LLC.

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

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

Chen, W. H., & Thissen, D. (1999). Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores. *British Journal of Mathematical and Statistical Psychology*, *52*, 19–37.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*, 474–497.

Ferrando, P. J., & Lorenzo-seva, U. (2001). Checking the appropriateness of item response theory models by predicting the distribution of observed scores: The program EO-fit. *Educational and Psychological Measurement*, *61*, 895–902.

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

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

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). Maximum marginal likelihood and expected a posteriori estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–288). Boston, MA: Kluwer.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, *26*, 195–211.

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

Ip, E. H. (2010a). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Ip, E. H. (2010b). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement*, *34*, 467–482.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*, 32–60.

Li, Y., & Rupp, A. A. (2011). Performance of the $S - X^2$ statistic for full-information bifactor models. *Educational and Psychological Measurement*, *71*, 986–1005.

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

Li, Z. & Cai, L. (2012). *Summed score based fit indices for testing latent variable distribution assumption in IRT. Paper presented at the 2012 International Meeting of the Psychometric Society*, Lincoln, NE.

Lord, F. M. (1953). The relation of test score to the latent trait underlying the test. *Educational and Psychological Measurement*, *13*, 517–548.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 453–461.

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

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*, 354–359.

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

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, 22–31.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696.

Rijmen, F. (2009). Efficient full information maximum likelihood estimation for multidimensional IRT models (Tech. Rep. No. RR-09-03). Educational Testing Service.

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

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*, 167–182.

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 253–292). Mahwah, NJ: Lawrence Erlbaum.

Ross, J. (1966). An empirical study of a logistic mental test model. *Psychometrika*, *31*, 325–340.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monographs No. 17)*. Richmond, VA: Psychometric Society.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, *37*, 41–57.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49.

Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL™ 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, *20*, 1497–1505.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Wu, E. J. C., & Bentler, P. M. (2011). *EQSIRT: A user-friendly IRT program (Computer software)*. Encino, CA: Multivariate Software.

Yung, Y. F., McLeod, L. D., & Thissen, D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.

Manuscript Received: 7 FEB 2013
Published Online Date: 19 SEP 2014