PSYCHOMETRIKA—VOL. 86, NO. 4, 973–993
DECEMBER 2021
https://doi.org/10.1007/s11336-021-09785-y

LORD–WINGERSKY ALGORITHM VERSION 2.5 WITH APPLICATIONS

SIJIA HUANG

INDIANA UNIVERSITY BLOOMINGTON

LI CAI

UNIVERSITY OF CALIFORNIA, LOS ANGELES (UCLA)

Item response theory scoring based on summed scores is employed frequently in the practice of educational and psychological measurement. Lord and Wingersky (Appl Psychol Meas 8(4):453–461, 1984) proposed a recursive algorithm to compute the summed score likelihood. Cai (Psychometrika 80(2):535–559, 2015) extended the original Lord–Wingersky algorithm to the case of two-tier multidimensional item factor models and called it Lord–Wingersky algorithm Version 2.0. The 2.0 algorithm utilizes dimension reduction to efficiently compute summed score likelihoods associated with the general dimensions in the model. The output of the algorithm is useful for various purposes, for example, scoring, scale alignment, and model fit checking. In the research reported here, a further extension to the Lord–Wingersky algorithm 2.0 is proposed. The new algorithm, which we call Lord–Wingersky algorithm Version 2.5, yields the summed score likelihoods for all latent variables in the model conditional on observed score combinations. The proposed algorithm is illustrated with empirical data for three potential application areas: (a) describing achievement growth using score combinations across adjacent grades, (b) identification of noteworthy subscores for reporting, and (c) detection of aberrant responses.

Key words: hierarchical item factor model, summed score, subscore, bifactor model.

1. Introduction

Generalizing the seminal Lord–Wingersky (1984) algorithm to other settings has been a regular topic in item response theory (IRT) research since its initial publication more than 35 years ago. Also well known in the Rasch modeling community (Andersen, 1972; Gustafsson, 1980), this simple recursive algorithm's wide-reaching impact in psychometrics is impressive to behold. For example, Hanson (1994), Thissen et al. (1995), as well as von Davier and Rost (1995), were among the first to expand the algorithm to polytomous IRT models. Chen and Thissen (1999) derived an item calibration algorithm based on summed scores. Thissen and Wainer's (2001) influential text on test scoring presented extensive methods for handling mixed-format tests, including an approach to handle score combinations (Rosa et al., 2001) that heavily influenced our thinking in the study reported here. Orlando et al. (2000) applied the Lord–Wingersky algorithm to illustrate summed score-based test linking, another area consistently of interest to psychometricians (e.g., Zeng & Kolen, 1995; Thissen et al., 2011). Orlando and Thissen (2000) proposed a solution to the item fit testing problem with a slight alteration of the original Lord–Wingersky algorithm. Li and Cai (2018) further extended the algorithm to create more accurate distributional approximations for test statistics sensitive to latent variable distributional assumptions. Stucky (2009), and independently Kim (2013), developed the weighted version of the algorithm wherein the item scores can take non-integer values.

Correspondence should be made to Li Cai, University of California, Los Angeles (UCLA), CRESST, 300 Charles E. Young Dr. North, GSEIS Building, Los Angeles, CA90095-1522, USA. Email: lcai@ucla.edu

Cai (2015) extended the algorithm to the case of hierarchical item factor models, specifically the two-tier model (Cai, 2010b). He named it Lord–Wingersky algorithm 2.0. In brief, a two-tier model consists of M primary latent dimensions ($\boldsymbol{\eta}$) and N specific latent dimensions (ξ_n, $n = 1, \ldots N$). The specific dimensions are independent conditional on the primary latent dimensions. Each item can load on at most one specific latent dimension, creating N non-overlapping *item clusters*. The item bifactor model (Gibbons & Hedeker, 1992), a member of the two-tier family (where $M = 1$), has experienced particular theoretical and empirical success recently (see Cai et al., 2011; Reise et al., 2007, 2018; Reise, 2012). In addition, the standard correlated-traits MIRT model (Reckase, 2009) and the testlet response theory model (Wainer et al., 2007) are constrained versions of the two-tier model. The two-tier structure permits the implementation of a dimension reduction technique (Rijmen, 2009) for computationally efficient maximum marginal likelihood parameter estimation with quadrature.

The dominating insight of Cai (2015) is that the non-overlapping item clusters are exchangeable conditional on the primary latent dimensions. In the original Lord–Wingersky algorithm, the items and their item scores are the basic building blocks. In the Lord–Wingersky algorithm 2.0, item clusters take the place of items and become the fungible units of model building and computation. Once again, dimension reduction can efficiently handle the numerical integration with quadrature. The algorithm yields summed score to scaled score conversions for the primary dimension(s), along with other associated statistical indices, with $(M + 1)$-fold integration regardless of the total number of factors in the model.

The present study extends the Lord–Wingersky algorithm 2.0. The new algorithm (Lord–Wingersky algorithm 2.5) uses patterns of item cluster summed scores instead of the overall summed score in Lord–Wingersky algorithm 2.0. Specifically, the item cluster summed score patterns are combinations of the observed score from one cluster and the summed score of the rest of the item clusters. It reduces to cluster score combinations when there are only two item clusters. It is worth noting here that the idea of using observed scores patterns to score individuals in unidimensional IRT is not new (e.g., Rosa et al., 2001). The algorithm proposed in this study generalizes this idea to scenarios where the underlying IRT models are hierarchical item factor models. Lord–Wingersky algorithm 2.5 leads to multidimensional posteriors of the primary latent dimension(s) with each specific latent dimension. The posterior probability of each observed score combination is a natural by-product. We illustrate applications of the proposed algorithm with three examples.

In the first example, we fit a longitudinal IRT model and use the Lord–Wingersky algorithm 2.5 to enhance the growth interpretation of score scales across adjacent grades in an operational large-scale English language proficiency assessment program, all without having to set a "vertical" scale. Second, the bivariate posteriors and score combination probabilities are used to facilitate the decision-making on subscore reporting. Finally, we construct posterior high-density region (HDR) for observed score combinations to help detect aberrant responses.

2. Lord–Wingersky Algorithm 2.0

We briefly review Cai's (2015) Lord–Wingersky algorithm 2.0 to establish notation.

With no loss of generality, consider a bifactor model with N specific latent dimensions, wherein each ξ_n is measured by I_n dichotomously scored items, and $n = 1, \ldots, N$. Let the prior (population) distribution of the general dimension η be denoted $h(\eta)$. To avoid notational clutter, instead of assuming conditional independence of the prior distributions of the specific dimensions $g(\xi_n|\eta)$ on η, we will assume, again with no loss of generality, fully independent specific dimensions. In other words, we shall write $g(\xi_n)$ as the prior of ξ_n. Define $T_i\ (1|\eta, \xi_n)$ as

the item response function of the ith item ($i = 1, \ldots I_n$) in cluster n, such that

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

where a_i^0 and a_i^n are the primary latent dimension and specific latent dimension item slopes, respectively, and c_i is the item intercept. The item parameters are assumed to be known and fixed, usually from a calibration study.

2.1. Stage I

In the first stage of Lord–Wingersky algorithm 2.0, for each item cluster, the within-cluster summed score likelihoods are accumulated over the latent space spanned by the primary latent dimension and the specific latent dimension. Let $P_i^n(x|\eta, \xi_n)$ denote the likelihood of summed score x after including the ith item in item cluster n in the recursive computation to be described below. Consider the nth item cluster, the algorithm initializes with the first item by starting the likelihood of summed score 0 $P_1^n\,(0|\eta, \xi_n)$ at the item response probabilities $T_1\,(0|\eta, \xi_n)$, and $P_1^n\,(1|\eta, \xi_n) = T_1\,(1|\eta, \xi_n)$. Then, the second item is added, resulting in three available summed scores: 0, 1, and 2. The corresponding summed score likelihoods after adding item 2 are:

$$\begin{aligned}
P_2^n\,(0|\eta, \xi_n) &= P_1^n\,(0|\eta, \xi_n)\,T_2\,(0|\eta, \xi_n)\,,\\
P_2^n\,(1|\eta, \xi_n) &= P_1^n\,(1|\eta, \xi_n)\,T_2\,(0|\eta, \xi_n) + P_1^n\,(0|\eta, \xi_n)\,T_2\,(1|\eta, \xi_n)\\
P_2^n\,(2|\eta, \xi_n) &= P_1^n\,(1|\eta, \xi_n)\,T_2\,(1|\eta, \xi_n)\,.
\end{aligned} \tag{2}$$

After this, each of the remaining items in item cluster n is included in the computation to form the desired within-cluster summed score likelihoods. Specifically, in step $i (2 < i \leq I_n)$ of the recursive algorithm, the ith item is added as follows:

$$\begin{aligned}
P_i^n\,(0|\eta, \xi_n) &= P_{i-1}^n\,(0|\eta, \xi_n)\,T_i\,(0|\eta, \xi_n)\\
P_i^n\,(x|\eta, \xi_n) &= P_{i-1}^n\,(x|\eta, \xi_n)\,T_i\,(0|\eta, \xi_n) + P_{i-1}^n\,(x-1|\eta, \xi_n)\,T_i\,(1|\eta, \xi_n)\\
P_i^n\,(i|\eta, \xi_n) &= P_i^n\,(i-1|\eta, \xi_n)\,T_i\,(1|\eta, \xi_n)\,.
\end{aligned} \tag{3}$$

The middle equation in (3) is repeated over values of x between 1 and $i - 1$.

To avoid notational clutter, let $P_n\,(s_n|\eta,\xi_n) = P_{I_n}^n\,(s_n|\eta, \xi_n)$ denote the likelihood associated with the within-cluster summed score $s_n = 0, \ldots, I_n$, after all I_n items in cluster n have been added according to the recursions defined in Eq. (3). At this point, an extra step is performed. The specific latent dimension, ξ_n, is integrated out, leaving the summed score likelihoods solely a function of the primary latent dimension, η. For simplicity, we can approximate this integral with rectangular quadrature:

$$P_n\,(s_n|\eta) = \int P_n\,(s_n|\eta,\xi_n)g\,(\xi_n)\,d\xi_n \approx \sum_{q=1}^{Q} P_n\left(s_n|\eta,Y_q\right)W_n(Y_q), \tag{4}$$

where Q is the number of quadrature points, Y_q the qth quadrature point, and $W_n(Y_q)$ is the corresponding quadrature weight, computed as normalized ordinates of $g\,(\xi_n)$.

2.2. Stage II

At the end of the first stage, available to us are N sets of within-cluster summed score likelihoods $\{P_n(s_n|\eta); s_n = 0, \ldots, I_n\}$, for $n = 1, \ldots, N$. These quantities depend only on the primary latent dimension η. Each item cluster can now be treated as if it were a polytomous item with $I_n + 1$ categories, and the "item scores" range from 0 to I_n.

Denote $L_n(s|\eta)$ as the likelihood of summed score s after adding item cluster n to the existing summed score likelihoods in the recursive computation described below. Let S_n be the maximum obtainable summed score after adding item cluster n. In our context when the items are all dichotomous, $S_n = \sum_{j=1}^{n} I_j$. Obviously S_N would be the maximum summed score. At this point, the standard Lord–Wingersky algorithm for polytomous items can be applied.

Let $L_1(s_1|\eta) = P_1(s_1|\eta), \forall s_1 = 0, \ldots, I_1$, for the purpose of initialization. Then in step $n (2 < n \leq N)$, the likelihoods $P_n(s_n|\eta)$ from item cluster n are added to the likelihoods from previous step to form the desired summed score likelihoods. For each possible summed score $0 \leq s \leq S_n$, we let

$$L_n(s|\eta) = \sum_{s_{n-1}=0}^{S_{n-1}} \sum_{s_n=0}^{I_n} L_{n-1}(s_{n-1}|\eta) P_n(s_n|\eta) \mathbf{1}_s(s_{n-1} + s_n), \tag{5}$$

where $\mathbf{1}_s(s_{n-1} + s_n)$ is an indicator function and takes the value of 1 if $s_{n-1} + s_n = s$ and 0 otherwise. Equation (5) essentially involves the booking keeping for a pair of scores s_{n-1} (from all item clusters added previously) and s_n (from the current item cluster) that adds up to the summed score s. When all N item clusters are included $L_N(s|\eta)$—or simply $L(s|\eta)$ to reduce clutter—contains the summed score likelihoods for the primary dimensions for $0 \leq s \leq S_N$.

2.3. Posterior Summaries

Recall that $h(\eta)$ is the prior distribution of the primary latent dimension. The normalized posterior of η associated with summed score s is

$$p(\eta|s) = \frac{L(s|\eta) h(\eta)}{p(s)} \tag{6}$$

where $p(s)$ is the (marginal) probability of summed score s:

$$p(s) = \int L(s|\eta) h(\eta) d\eta \approx \sum_{q=1}^{Q} L\left(s|Y_q\right) W\left(Y_q\right), \tag{7}$$

and Q rectangular quadrature points X_q are used to approximate the posterior, with $W\left(X_q\right)$ the normalized ordinates of $h(\eta)$. The posterior mean $E(\eta|s)$ and posterior variance $Var(\eta|s) = E\left(\eta^2|s\right) - E^2(\eta|s)$ are useful summaries, where

$$\begin{aligned} E(\eta|s) &= \frac{1}{p(s)} \int \eta L(s|\eta) h(\eta) d\eta \approx \frac{1}{p(s)} \sum_{q=1}^{Q} X_q L\left(s|X_q\right) W\left(X_q\right), \\ E\left(\eta^2|s\right) &= \frac{1}{p(s)} \int \eta^2 L(s|\eta) h(\eta) d\eta \approx \frac{1}{p(s)} \sum_{q=1}^{Q} X_q^2 L\left(s|X_q\right) W\left(X_q\right). \end{aligned} \tag{8}$$

A normal approximation of the posterior based on the posterior mean and variance often works quite well even when the number of items is moderate. The posterior mean can be used as the summed score-based IRT scaled score estimate and the posterior variance as the error variance estimate for the scaled score. The marginal probability $p(s)$ itself can be useful either as a model-based (pre-operational) estimated of the expected summed score group probability or as an aid in IRT model fit checking.

3. Lord–Wingersky Algorithm 2.5

3.1. General Approach

Recall the bifactor model with N specific latent dimensions defined in Sect. 2. Each of the N item clusters includes I_n items. The Lord–Wingersky algorithm 2.0 is focused on obtaining the posterior distribution of the primary dimension η, conditioned on the overall summed score. The specific latent dimensions are integrated out at the end of Stage I (see Sect. 2.1). In the proposed algorithm, we obtain bivariate posteriors of the primary latent dimension η and the specific latent dimension ξ_n.

Instead of the overall summed score, each bivariate posterior is conditioned on a pair of scores. We continue to use s_n to denote the summed scores from item cluster n, where $s_n = 0, \ldots, I_n$, and introduce here new notation for the *rest score* $s_{(n)}$, i.e., the summed score from all clusters except item cluster n. Let $S_{(n)} = \sum_{j=1, j\neq n}^{N} I_j$ be the maximum summed score from the rest of the item clusters so $s_{(n)} = 0, \ldots, S_{(n)}$. Cai (2015) in fact alluded to the possibility of using the summed vs. rest score combination $(s_n, s_{(n)})$, but stopped shy of actually computing the bivariate posterior, as we now outline below.

3.2. Stage I

In the first stage, for each item cluster, the within-cluster summed score likelihoods are accumulated over the space spanned by η and $\xi_n, n = 1, \ldots N$, just as in Lord–Wingersky algorithm 2.0. At the end of this stage, we retain and store the likelihoods for the primary dimension $\{P_n(s_n|\eta); s_n = 0, \ldots, I_n\}, \forall n = 1, \ldots, N$. The critical added requirement is that we also retain and store all the within-cluster summed score likelihoods $\{P_n(s_n|\eta, \xi_n); s_n = 0, \ldots, I_n\}, \forall n = 1, \ldots, N$. In a quadrature representation of the likelihoods, at most $Q \times Q$ floating point values are stored per cluster, per score, if Q quadrature points per dimension are used.

3.3. Stage II

We will now cycle through the item clusters to compute the desired bivariate posteriors. In general, for item cluster n, we wish to construct bivariate posteriors for η and ξ_n. Recall that the other item clusters do not depend on ξ_n, so we proceed by treating the cluster summed score likelihood values $P_1(s_1|\eta), \ldots, P_{n-1}(s_{n-1}|\eta), P_{n+1}(s_{n+1}|\eta), \ldots, P_N(s_N|\eta)$ as though they were polytomous items that depend on η. The standard Lord–Wingersky algorithm can now be applied readily to produce item cluster n's rest score likelihoods $R_n\left(s_{(n)}|\eta\right)$, for $s_{(n)} = 0, \ldots, S_{(n)}$. In other words, the recursions work in exactly the same manner as Sect. 2.2, except that we omit the likelihood contributions from $P_n(s_n|\eta)$.

The rest score likelihoods $R_n\left(s_{(n)}|\eta\right)$ are then combined with the summed score likelihoods from item cluster n, $P_n(s_n|\eta, \xi_n), s = 0, \ldots, I_n$, as well as the prior distributions for η and ξ_n, to yield the bivariate posterior distributions of η and ξ_n associated with the summed vs. rest score combination $(s_n, s_{(n)})$:

TABLE 1.
Item parameters of the six-item scale

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

$$p\left(\eta, \xi_n | s_n, s_{(n)}\right) = \frac{P_n\left(s_n|\eta, \xi_n\right) R_n\left(s_{(n)}|\eta\right) g\left(\xi_n\right) h\left(\eta\right)}{p\left(s_n, s_{(n)}\right)}, \tag{9}$$

where the marginal probability $p\left(s_n, s_{(n)}\right)$ is

$$p\left(s_n, s_{(n)}\right) = \iint P_n\left(s_n|\eta, \xi_n\right) R_n\left(s_{(n)}|\eta\right) g\left(\xi_n\right) h\left(\eta\right) d\xi_n d\eta. \tag{10}$$

Again, the posterior above can be easily approximated with rectangular quadrature:

$$p\left(s_n, s_{(n)}\right) \approx \sum_{r=1}^{Q}\sum_{q=1}^{Q} P_n\left(s_n|X_r, Y_q\right) R_n\left(s_{(n)}|X_r\right) W_n\left(Y_q\right) W(X_r). \tag{11}$$

Aside from the marginal probability in Eq. (10), other useful summaries of the posterior distribution include the mean vector $\boldsymbol{\mu}$ the covariance matrix $\boldsymbol{\Sigma}$, which facilitate a bivariate normal approximation to the posterior that can be quite effective in practice, as we shall demonstrate later. The marginal posterior means $\mu_0 = E\left(\eta|s_n, s_{(n)}\right)$ and $\mu_n = E\left(\xi_n|s_n, s_{(n)}\right)$, and the error variances and covariance $\sigma_{00} = Var\left(\eta|s_n, s_{(n)}\right)$, $\sigma_{0n} = Cov\left(\eta, \xi_n|s_n, s_{(n)}\right)$, and $\sigma_{nn} = Var\left(\xi_n|s_n, s_{(n)}\right)$ provide reasonable point estimates and error (co)variance estimates for all the latent variables in the model. These means and covariance matrix elements can be approximated with quadrature along similar lines as Eq. (11).

3.4. An Illustrative Example

Consider the same hypothetical six-item scale with bifactor structure as discussed in Cai (2015). These six dichotomous items form three item clusters, each consisting of two items. Priors of all four latent dimensions (one primary latent dimension and three specific latent dimensions) are assumed independent and standard normal. Table 1 shows the item parameters and the factor pattern.

For each dimension, $Q = 5$ equally spaced quadrature points at $-2, -1, 0, 1$, and 2 are used for demonstrate purposes only (practical usage of the algorithm requires much larger values of Q). Thus, a 5×5 grid is formed as the direct product of η and each of the three specific latent dimensions when appropriate. Summed score likelihoods are evaluated over these grid points.

Tables 2 and 3 show the first stage of the Lord–Wingersky algorithm 2.5 to compute the bivariate posterior of η and ξ_1. In Table 2, the within-cluster summed score likelihoods of the three

TABLE 2.
Accumulating within-in summed score likelihoods for item cluster 1, 2, and 3

Quadrature grid for (η, ξ_1)									
Initializing with item 1 in cluster 1, having two available summed scores									
η	-2	-2	-2	$\cdots$	0	$\cdots$	2	2	2
ξ_1	-2	-1	0	$\cdots$	0	$\cdots$	0	1	2
$P_1^1(0\|\eta,\xi_1) = T_1(0\|\eta,\xi_1)$	.996	.988	.968	$\cdots$	.731	$\cdots$	.198	.083	.032
$P_1^1(1\|\eta,\xi_1) = T_1(1\|\eta,\xi_1)$	.004	.012	.032	$\cdots$	.269	$\cdots$	.802	.917	.968
Adding item 2 to the computation									
$P_2^1(0\|\eta,\xi_1) =$ $P_1^1(0\|\eta,\xi_1)\,T_2(0\|\eta,\xi_1)$	.989	.970	.922	$\cdots$	.472	$\cdots$	.028	.005	.001
$P_2^1(1\|\eta,\xi_1) =$ $P_1^1(0\|\eta,\xi_1)\,T_2(1\|\eta,\xi_1) +$ $P_1^1(1\|\eta,\xi_1)\,T_2(0\|\eta,\xi_1)$	.011	.030	.077	$\cdots$	.433	$\cdots$	.284	.131	.053
$P_2^1(2\|\eta,\xi_1) =$ $P_1^1(1\|\eta,\xi_1)\,T_2(1\|\eta,\xi_1)$	.000	.000	.002	$\cdots$	.095	$\cdots$	.688	.864	.947
Quadrature grid for (η, ξ_2)									
Initializing with item 1 in cluster 2, having two available summed scores									
η	-2	-2	-2	$\cdots$	0	$\cdots$	2	2	2
ξ_2	-2	-1	0	$\cdots$	0	$\cdots$	0	1	2
$P_1^2(0\|\eta,\xi_2) = T_1(0\|\eta,\xi_2)$	.978	.953	.900	$\cdots$	.550	$\cdots$	.142	.069	.032
$P_1^2(1\|\eta,\xi_2) = T_1(1\|\eta,\xi_2)$	.022	.047	.100	$\cdots$	.450	$\cdots$	.858	.931	.968
Adding item 2 to the computation									
$P_2^2(0\|\eta,\xi_2) =$ $P_1^2(0\|\eta,\xi_2)\,T_2(0\|\eta,\xi_2)$	.947	.887	.773	$\cdots$	.248	$\cdots$	.014	.003	.001
$P_2^2(1\|\eta,\xi_2) =$ $P_1^2(0\|\eta,\xi_2)\,T_2(1\|\eta,\xi_2) +$ $P_1^2(1\|\eta,\xi_2)\,T_2(0\|\eta,\xi_2)$	.053	.110	.213	$\cdots$	.505	$\cdots$	.213	.110	.053
$P_2^2(2\|\eta,\xi_2) =$ $P_1^2(1\|\eta,\xi_2)\,T_2(1\|\eta,\xi_2)$	.001	.003	.014	$\cdots$	.248	$\cdots$	.773	.887	.947
Quadrature grid for (η, ξ_3)									
Initializing with item 1 in cluster 3, having two available summed scores									
η	-2	-2	-2	$\cdots$	0	$\cdots$	2	2	2
ξ_3	-2	-1	0	$\cdots$	0	$\cdots$	0	1	2
$P_1^3(0\|\eta,\xi_3) = T_1(0\|\eta,\xi_3)$	.968	.900	.731	$\cdots$	.354	$\cdots$	.100	.032	.010
$P_1^3(1\|\eta,\xi_3) = T_1(1\|\eta,\xi_3)$	.032	.100	.269	$\cdots$	.646	$\cdots$	.900	.968	.990
Adding item 2 to the computation									
$P_2^3(0\|\eta,\xi_3) =$ $P_1^3(0\|\eta,\xi_3)\,T_2(0\|\eta,\xi_3)$	.922	.773	.472	$\cdots$	.095	$\cdots$	.007	.001	.000
$P_2^3(1\|\eta,\xi_3) =$ $P_1^3(0\|\eta,\xi_3)\,T_2(1\|\eta,\xi_3) +$ $P_1^3(1\|\eta,\xi_3)\,T_2(0\|\eta,\xi_3)$	.077	.213	.433	$\cdots$	.433	$\cdots$	.155	.053	.017
$P_2^3(2\|\eta,\xi_3) =$ $P_1^3(1\|\eta,\xi_3)\,T_2(1\|\eta,\xi_3)$	.002	.014	.095	$\cdots$	.472	$\cdots$	.838	.947	.983

TABLE 3.
Integrating the specific dimensions ξ_2 and ξ_3 out of the summed score likelihoods

	Quadrature grid for (η, ξ_2)								
η	-2	-2	-2	...	0	...	2	2	2
ξ_2	-2	-1	0	...	0	...	0	1	2
$W_2(\xi_2)$	.054	.244	.403	...	.403	...	.403	.244	.054
Multiplying cluster 2's summed score likelihoods by $W_2(\xi_2)$									
$P_2(0\|\eta,\xi_2)\,W_2(\xi_2) = P_2^2(0\|\eta,\xi_2)\,W_2(\xi_2)$	052	.217	.311	...	.100	...	.006	.001	.000
$P_2(1\|\eta,\xi_2)\,W_2(\xi_2) = P_2^2(1\|\eta,\xi_2)\,W_2(\xi_2)$	.003	.027	.086	...	.203	...	.086	.027	.003
$P_2(2\|\eta,\xi_2)\,W_2(\xi_2) = P_2^2(2\|\eta,\xi_2)\,W_2(\xi_2)$	.000	.001	.006	...	.100	...	.311	.217	.052

	Quadrature grid for η				
	-2	-1	0	1	2
Leaving cluster'2 summed score as a function of η, by integrating out ξ_2 (summing over X_q)					
$P_2(0\|\eta) = \sum_{\xi_2} P_2(0\|\eta,\xi_2)\,W_2(\xi_2)$	.742	.519	.277	.106	.028
$P_2(1\|\eta) = \sum_{\xi_2} P_2(0\|\eta,\xi_2)\,W_2(\xi_2)$	.230	.375	.446	.375	.230
$P_2(2\|\eta) = \sum_{\xi_2} P_2(0\|\eta,\xi_2)\,W_2(\xi_2)$	.028	.106	.277	.519	.742

	Quadrature grid for (η, ξ_3)								
η	-2	-2	-2	$\cdots$	0	$\cdots$	2	2	2
ξ_3	-2	-1	0	$\cdots$	0	$\cdots$	0	1	2
$W_3(\xi_3)$	.054	.244	.403	$\cdots$	.403	$\cdots$	.403	.244	.054
Multiplying cluster 3's summed score likelihoods by $W_3(\xi_3)$									
$P_3(0\|\eta,\xi_3)\,W_3(\xi_3) = P_2^3(0\|\eta,\xi_3)\,W_3(\xi_3)$	.050	.189	.190	$\cdots$	.038	$\cdots$	.003	.000	.000
$P_3(1\|\eta,\xi_3)\,W_3(\xi_3) = P_2^3(1\|\eta,\xi_3)\,W_3(\xi_3)$	.004	.052	.174	$\cdots$	.174	$\cdots$	.062	.013	.001
$P_3(2\|\eta,\xi_3)\,W_3(\xi_3) = P_2^3(2\|\eta,\xi_3)\,W_3(\xi_3)$	.000	.003	.038	$\cdots$	.190	$\cdots$	.337	.231	.054

	Quadrature grid for η				
	-2	-1	0	1	2
Leaving cluster'3 summed score as a function of η, by integrating out ξ_3 (summing over ξ_3)					
$P_3(0\|\eta) = \sum_{\xi_3} P_3(0\|\eta,\xi_3)\,W_3(\xi_3)$	.469	.302	.166	.077	.029
$P_3(1\|\eta) = \sum_{\xi_3} P_3(1\|\eta,\xi_3)\,W_3(\xi_3)$	.364	.396	.364	.285	.192
$P_3(2\|\eta) = \sum_{\xi_3} P_3(2\|\eta,\xi_3)\,W_3(\xi_3)$	.166	.302	.469	.638	.779

item clusters are computed. In Table 3, specific dimensions ξ_2 and ξ_3 are integrated out, leaving the observed summed score of item cluster 2 and 3 as a function of η. Table 4 shows the second stage of the algorithm, where item clusters 2 and 3 are treated as polytomous items (both with 3 categories), while the rest score likelihoods for item cluster 1 are calculated. Table 5 shows how the bivariate posteriors associated with each summed vs. rest score pattern are computed. Summaries of the bivariate normal approximations of posteriors associated with the score combinations are shown in Table 6.

Figure 1 shows the equal probability contours of the bivariate normal approximated posteriors for five score combinations (with the mean vectors and covariance matrices in Table 6). Each contour includes 25% of the volume under the posterior density. The five score combinations

TABLE 4.
Forming the rest score likelihoods (summed score likelihoods for item clusters 2 and 3)

	Quadrature Grid for η				
	-2	-1	0	1	2
Initializing with cluster 2, having 3 available summed scores					
$L_2(0\|\eta) = P_2(0\|\eta)$	.742	.519	.277	.106	.028
$L_2(1\|\eta) = P_2(1\|\eta)$	.230	.375	.446	.375	.230
$L_2(2\|\eta) = P_2(2\|\eta)$	.028	.106	.277	.519	.742
Adding cluster 3' summed scores to item cluster 1's rest score likelihoods					
$R_1(0\|\eta) = L_3(0\|\eta) = L_2(0\|\eta)P_3(0\|\eta)$	.348	.157	.046	.008	.001
$R_1(1\|\eta) = L_3(1\|\eta) = L_2(0\|\eta)\,P_3(1\|\eta) + L_2(1\|\eta)\,P_3(0\|\eta)$	.378	.319	.175	.059	.012
$R_1(2\|\eta) = L_3(2\|\eta) = L_2(0\|\eta)\,P_3(2\|\eta) + L_2(1\|\eta)\,P_3(1\|\eta) + L_2(2\|\eta)P_3(0\|\eta)$	.220	.337	.339	.214	.088
$R_1(3\|\eta) = L_3(3\|\eta) = L_2(1\|\eta)\,P_3(2\|\eta) + L_2(2\|\eta)P_3(1\|\eta)$	.049	.155	.310	.387	.321
$R_1(4\|\eta) = L_3(4\|\eta) = L_2(2\|\eta)P_3(2\|\eta)$	.005	.032	.130	.331	.578

TABLE 5.
Forming posteriors for score combinations related to item cluster 1

	Quadrature grid for (η, ξ_1)								
η	-2	-2	-2	$\cdots$	0	$\cdots$	2	2	2
ξ_1	-2	-1	0	$\cdots$	0	$\cdots$	0	1	2
$W(\eta)$	.054	.054	.054	$\cdots$	.403	$\cdots$	.054	.054	.054
$W_1(\xi_1)$	.054	.244	.403	$\cdots$	.403	$\cdots$	.403	.244	.054
$p(0,0\|\eta,\xi_1) \propto P_1(0\|\eta,\xi_1)R_1(0\|\eta)\,W(\eta)W_1(\xi_1)$	.0010	.0045	.0070	$\cdots$	.0035	$\cdots$	.0000	.0000	.0000
$p(1,0\|\eta,\xi_1) \propto P_1(1\|\eta,\xi_1)R_1(0\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0001	.0006	$\cdots$	.0032	$\cdots$	.0000	.0000	.0000
$p(2,0\|\eta,\xi_1) \propto P_1(2\|\eta,\xi_1)R_1(0\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0000	$\cdots$	.0007	$\cdots$	.0000	.0000	.0000
$p(0,1\|\eta,\xi_1) \propto P_1(0\|\eta,\xi_1)R_1(1\|\eta)\,W(\eta)W_1(\xi_1)$	.0011	.0049	.0077	$\cdots$	.0134	$\cdots$	.0000	.0000	.0000
$p(1,1\|\eta,\xi_1) \propto P_1(1\|\eta,\xi_1)R_1(1\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0001	.0006	$\cdots$	.0123	$\cdots$	.0001	.0000	.0000
$p(2,1\|\eta,\xi_1) \propto P_1(2\|\eta,\xi_1)R_1(1\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0000	$\cdots$	.0027	$\cdots$	.0002	.0001	.0000
$p(0,2\|\eta,\xi_1) \propto P_1(0\|\eta,\xi_1)R_1(2\|\eta)\,W(\eta)W_1(\xi_1)$	.0006	.0028	.0045	$\cdots$	.0259	$\cdots$	.0001	.0000	.0000
$p(1,2\|\eta,\xi_1) \propto P_1(1\|\eta,\xi_1)R_1(2\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0001	.0004	$\cdots$	.0237	$\cdots$	.0005	.0002	.0000
$p(2,2\|\eta,\xi_1) \propto P_1(2\|\eta,\xi_1)R_1(2\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0000	$\cdots$	.0052	$\cdots$	.0013	.0010	.0002
$p(0,3\|\eta,\xi_1) \propto P_1(0\|\eta,\xi_1)R_1(3\|\eta)\,W(\eta)W_1(\xi_1)$	.0001	.0006	.0010	$\cdots$	.0237	$\cdots$	.0002	.0000	.0000
$p(1,3\|\eta,\xi_1) \propto P_1(1\|\eta,\xi_1)R_1(3\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0001	$\cdots$	.0218	$\cdots$	.0020	.0006	.0001
$p(2,3\|\eta,\xi_1) \propto P_1(2\|\eta,\xi_1)R_1(3\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0000	$\cdots$	.0048	$\cdots$	.0049	.0037	.0009
$p(0,4\|\eta,\xi_1) \propto P_1(0\|\eta,\xi_1)R_1(4\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0001	.0001	$\cdots$	.0100	$\cdots$	.0004	.0000	.0000
$p(1,4\|\eta,\xi_1) \propto P_1(1\|\eta,\xi_1)R_1(4\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0000	$\cdots$	.0091	$\cdots$	.0036	.0010	.0001
$p(2,4\|\eta,\xi_1) \propto P_1(2\|\eta,\xi_1)R_1(4\|\eta)\,W(\eta)W_1(\xi_1)$	.0000	.0000	.0000	$\cdots$	.0020	$\cdots$	.0087	.0066	.0016

$(s_1, s_{(1)})$ are (0, 0), (0, 2),(0,4),(1,2) and (2,2), and the corresponding posterior means of ξ_1 are: $-.232$, $-.370$, $-.573$, .212, and .732, and those of η are: -1.136, $-.477$, .302, .025, and .492.

Examining the means and variances reveals some interesting patterns. First, all the posterior covariances between the primary and specific dimension are negative, even when they are *a priori* uncorrelated. Second, when the rest score remains the same (2), the specific dimension score increases from $-.370$ to .212 and .732 as the summed score for item cluster 1 increases from 0 to 1 and 2, which is to be expected. Third, because the total summed score for the combination (0,4) is 4, and for the combination (1, 2) is 3, we intuit that the second combination should imply

TABLE 6.
Summaries of posteriors associated with each score combination

	Prob.	μ_0	σ_{00}	μ_1	σ_{11}	σ_{01}
$s_1=0, s_{(1)}=0$	.054	−1.136	.488	−.232	.815	−.091
$s_1=1, s_{(1)}=0$	.019	−.640	.528	.413	.756	−.148
$s_1=2, s_{(1)}=0$	.005	−.168	.513	.930	.640	−.150
$s_1=0, s_{(1)}=1$	.111	−.812	.540	−.296	.796	−.113
$s_1=1, s_{(1)}=1$	.053	−.304	.533	.315	.754	−.162
$s_1=2, s_{(1)}=1$	.019	.162	.519	.832	.664	−.156
$s_1=0, s_{(1)}=2$	.146	−.477	.560	−.370	.775	−.131
$s_1=1, s_{(1)}=2$	.096	.025	.536	.212	.753	−.172
$s_1=2, s_{(1)}=2$	.046	.492	.527	.732	.688	−.159
$s_1=0, s_{(1)}=3$	.110	−.091	.552	−.466	.750	−.144
$s_1=1, s_{(1)}=3$	.101	.392	.531	.092	.752	−.177
$s_1=2, s_{(1)}=3$	.067	.850	.511	.624	.711	−.151
$s_1=0, s_{(1)}=4$	.050	.302	.545	−.573	.725	−.155
$s_1=1, s_{(1)}=4$	.064	.771	.519	−.036	.751	−.175
$s_1=2, s_{(1)}=4$	.059	1.205	.456	.521	.731	−.131

a lower primary dimensional score (.302 vs. .025), as expected. Fourth, holding item cluster 1 score constant (0), the primary dimension score is increasing (from −1.136 to −.477 and to .302) as the rest score moves from 0 to 1 and 2, again as expected. Fifth, the posterior means of the specific dimension, on the other hand, decreases from −.232, to −.370, and ultimately to −.573, consistent with the negative posterior correlations between the primary and specific dimensions. In this case, the posterior variance shrinks from .815 to .775 to .725, indicating increasing certainty in the specific dimension score. Finally, the posterior variance for the general dimension is .536 for the score combination (1, 2). This is slightly smaller than the reported posterior variance (.55) of the general dimension in Cai (2015) for same total score (3), because the latter variance is conditional on the total summed score alone, a further reduction of the observed data.

3.5. The More General Case

The algorithm generalizes naturally when the number of general dimensions exceeds 1 ($M > 1$). In this case, the single set of rectangular quadrature points Y_q that cover the latent variable space of $\boldsymbol{\eta}$ becomes a direct product of M sets of quadrature points. The marginal posterior means $\boldsymbol{\mu}_0 = E\left(\boldsymbol{\eta}|s_n, s_{(n)}\right)$ is now a $M \times 1$ vector, the primary dimension error covariance matrix $\boldsymbol{\Sigma}_{00} = Var\left(\boldsymbol{\eta}|s_n, s_{(n)}\right)$ is $M \times M$, and the covariance terms in $\boldsymbol{\sigma}_{0n} = Cov\left(\boldsymbol{\eta}, \xi_n|s_n, s_{(n)}\right)$ become a $M \times 1$ vector.

It is worth noting if there are more than two item clusters, the estimates of $\boldsymbol{\eta}$ will change depending on which item cluster is treated as the focal cluster (e.g., cluster 1 vs. the rest, or cluster 2 vs. the rest, etc.). This phenomenon is analogous to the difference between response pattern-based scaled scores and summed scores-based scaled scores in unidimensional IRT models that do not assume equal item slope parameters. Items with varying slope parameters are not equally discriminating, and different response patterns with the same summed score will necessarily lead to different scaled score estimates. This is well understood. In hierarchical item factor models, item clusters take the place of items. The item clusters may have different difficulty and discriminability as far as $\boldsymbol{\eta}$ is concerned, and therefore different ways of decomposing the total summed score will lead to different $\boldsymbol{\eta}$ estimates. For a reader whose only concern is scoring for $\boldsymbol{\eta}$, Cai's

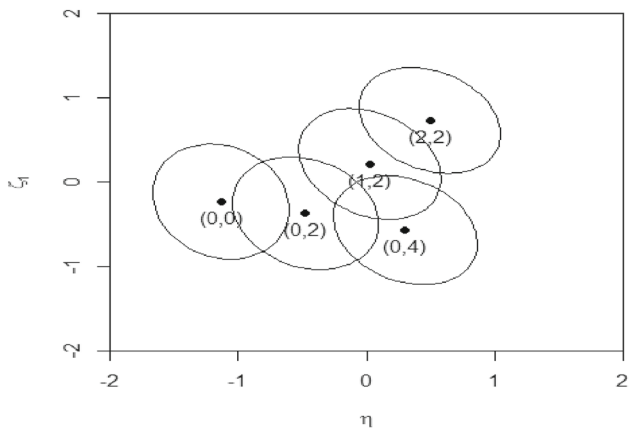

FIGURE 1.
Normal approximations to the posteriors of η and ξ_1 for five score combinations

(2015) algorithm may be simpler and more easily interpretable, though of course, cluster score combinations do condition on varied patterns of responses and contain more information than the total score.

4. Illustrative Applications of Lord–Wingersky Algorithm 2.5

4.1. Growth Interpretation of Observed Score Combinations

ELPA21 is a multi-state assessment program that provides measures of English language proficiency of English Learners (ELs) in K-12 educational systems in the participating states. It measures ELs' proficiencies in four language domains—reading, listening, writing, and speaking—from kindergarten to high school (i.e., kindergarten, grade 1, grade band 2–3, grade band 4–5, grade band 6–8, and grade band 9–12). Cut scores were established from standard setting studies in each domain and grade band so that students are classified as *emerging*, *progressing*, or *proficient*. Parameters of items in tests of different grade bands were calibrated separately (CRESST, 2017). Thus, the latent ability scales of two adjacent grade bands are different and the scale scores of tests of different grade bands cannot be compared directly.

A convenient and transparent way to report students' growth is desired, but ELPA21 took the position that vertical scaling, a technique popular in statewide accountability testing, should not employed. The major reason for the choice was that language learning and proficiency development change rapidly over time, especially in the early stage (e.g., PK to 2), potentially shifting the construct being measured considerably. Measures of proficiency such as ELPA21 test scores probably should not be compared directly (or forced on the same scale) for a PK English Learner vs. an English Learner in upper elementary. Hansen and Monroe (2018) provide additional discussions of this topic. Here, we explore the possible use of observed score combinations to describe student growth through the application of Lord–Wingersky algorithm 2.5.

Consider two fixed-form tests of in adjacent grade bands—one in the lower-grade band and one in the upper-grade band. The example here is the listening tests from grade bands 4–5 and 6–8. The lower-grade band test consists of 24 dichotomous items, and upper-grade band test includes 30 dichotomous items. A random sample of 300 students who took both tests was used to estimate the population distribution, while item parameters were held at the pre-calibrated values (see Table 7). Items in each test form are thought to load on either the lower- or upper-grade band

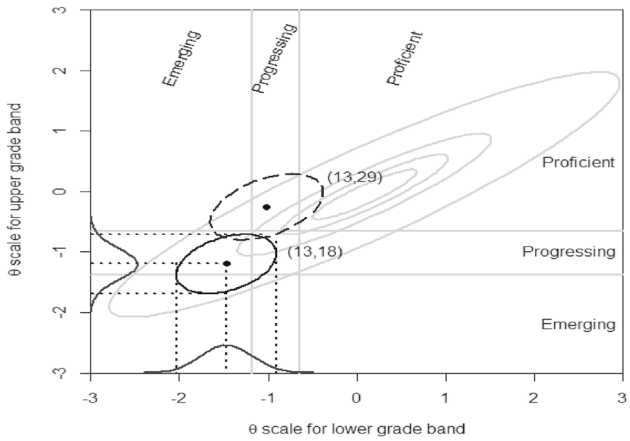

FIGURE 2.
MIRT model to aid ELPA21 growth score interpretation

listening proficiency latent variables, depending on which form they come from. The two latent variables are correlated, resulting in a classical correlated-traits MIRT model for longitudinal data. The estimated population means of the lower- and upper-grade band latent proficiency variables is .09 (SE = .08) and − .05 (SE = .06), respectively. Estimated population variances of the two grade bands are 1.25 (SE = .15) and .62 (SE = .07), and their covariance is .80 (SE = .09), yielding a correlation of .91.

The x-axis of Fig. 2 is the latent proficiency scale of the lower-grade band, and the y-axis is the latent proficiency scale of the upper-grade band. The estimated (bivariate normal) population distribution is overlaid in light gray. Four cut scores—two for the lower-grade band (− 1.1875 and − 0.65) and two for the upper-grade band (− 1.375 and − 0.65)—divide the space into nine regions. Bivariate normal approximations of posteriors associated with two score combinations are plotted. The two-dimensional MIRT model is treated as a two-tier model with empty specific latent dimensions (no item loading) so that the score combination posteriors can be computed via the Lord–Wingersky algorithm 2.5 implemented in flexMIRT® (Cai, 2017) without additional programming. This is analogous to Cai's (2015) Example 4.2 that replicates more specialized score combination computations, wherein the bifactor model was strictly not needed. The classification of *emerging*, *progressing*, or *proficient* are made out of the volume of the marginal posterior distribution that falls between the cut scores. For example, the probabilities of *emerging*, *progressing*, or *proficient* of students with observed score combination (13, 18) are .84, .16, and 0 in the lower-grade band and .24, .75, and .01 in the upper-grade band. We may then communicate clearly to the users of the score reports that this particular combination of 13 (out of 24) on the lower-grade band test and 18 (out of 30) on the upper-grade band test indicates an improvement from *emerging* to *progressing*. In a similar fashion, the score combination (13, 29) represents an improvement from *progressing* to *proficient*.

The probabilities of each of the combinations are also natural by-products of our recursive algorithm. Among students who received a score of 13 on the lower-grade band test (expected to be roughly 1.89% of the student population, based on the model), a score of 24 on the upper-grade band test places the student at the 74% percentile, which is akin to a student growth percentile (SGP; Betebenner, 2009) but entirely based on observed scores. In addition, although not pursued here, the Lord–Wingersky algorithm 2.5, coupled with the calibrated projection method (Thissen et al., 2011), can be applied to predict scores of the upper grade-band test based on the lower-

TABLE 7.
Item parameters of the ELPA 21 test forms in two consecutive years

Grade band	Item ID	Intercept	Slopes			
			η_1	η_2	ξ_1	ξ_2
Lower	1	3.26	1.43	0.00	0.00	
Lower	2	2.95	1.53	0.00	0.00	
Lower	3	1.10	0.46	0.00	0.00	
Lower	4	2.85	1.88	0.00	0.00	
Lower	5	1.95	1.51	0.00	0.00	
Lower	6	1.59	1.10	0.00	0.00	
Lower	7	2.82	1.50	0.00	0.00	
Lower	8	4.02	1.64	0.00	0.00	
Lower	9	0.18	0.29	0.00	0.00	
Lower	10	2.08	1.27	0.00	0.00	
Lower	11	2.24	1.28	0.00	0.00	
Lower	12	1.70	0.95	0.00	0.00	
Lower	13	4.34	1.71	0.00	0.00	
Lower	14	2.80	1.52	0.00	0.00	
Lower	15	3.77	2.10	0.00	0.00	
Lower	16	2.96	1.80	0.00	0.00	
Lower	17	3.33	1.79	0.00	0.00	
Lower	18	0.33	0.76	0.00	0.00	
Lower	19	−0.95	0.57	0.00	0.00	
Lower	20	2.18	1.46	0.00	0.00	
Lower	21	1.79	1.20	0.00	0.00	
Lower	22	1.78	1.57	0.00	0.00	
Lower	23	2.49	1.65	0.00	0.00	
Lower	24	1.38	1.01	0.00	0.00	
Upper	1	1.60	0.00	1.48		0.00
Upper	2	0.98	0.00	1.54		0.00
Upper	3	2.34	0.00	1.73		0.00
Upper	4	1.65	0.00	1.42		0.00
Upper	5	2.20	0.00	1.64		0.00
Upper	6	0.89	0.00	1.10		0.00
Upper	7	2.94	0.00	2.03		0.00
Upper	8	2.70	0.00	1.32		0.00
Upper	9	5.40	0.00	2.64		0.00
Upper	10	3.51	0.00	2.24		0.00
Upper	11	5.40	0.00	2.73		0.00
Upper	12	4.34	0.00	2.16		0.00
Upper	13	4.09	0.00	2.14		0.00
Upper	14	6.03	0.00	2.04		0.00
Upper	15	5.76	0.00	2.95		0.00
Upper	16	4.94	0.00	2.10		0.00
Upper	17	4.71	0.00	2.56		0.00
Upper	18	7.92	0.00	2.95		0.00
Upper	19	2.67	0.00	1.66		0.00
Upper	20	2.45	0.00	1.81		0.00
Upper	21	0.66	0.00	1.49		0.00

TABLE 7.
continued

Grade band	Item ID	Intercept	Slopes			
			η_1	η_2	ξ_1	ξ_2
Upper	22	2.14	0.00	1.74		0.00
Upper	23	1.51	0.00	1.81		0.00
Upper	24	1.93	0.00	1.39		0.00
Upper	25	2.51	0.00	1.90		0.00
Upper	26	2.72	0.00	1.86		0.00
Upper	27	3.85	0.00	2.31		0.00
Upper	28	0.35	0.00	0.73		0.00
Upper	29	2.28	0.00	1.65		0.00
Upper	30	1.39	0.00	1.26		0.00

grade band test scores. In sum, Lord–Wingersky algorithm 2.5 serves as a useful tool to facilitate reporting of student growth in the multi-state EL assessment program.

4.2. Facilitating Subscore Reporting

Educational and psychological assessments usually consist of several item clusters, yielding the so-called subscores. Within the IRT framework, several subscoring approaches, including the bifactor model approach and the correlated-traits MIRT model approach, are available. Subscore reporting is another recurrent topic in recent psychometrics literature (e.g., Sinharay et al., 2007; Haberman, 2008; Haberman et al. 2009; Feinberg & Wainer, 2014) because of the increasing demand for more detailed information about individuals. Two issues must be considered when deciding whether to report subscores obtained through a bifactor model (i.e., the ξ_n estimates) in addition to the overall score (i.e., the η estimate). The first question—if these subscores are reliable enough—is the easier one to address within the IRT framework. Here we focus on the second question—whether the information the subscores provide is distinct enough from the overall score.

We believe that if a subscore is considered to be surprising given an individual's overall score (i.e., if the ξ_n estimate cannot be well predicted by the η estimate), it should be reported for it is adding information. This is similar in spirit to Feinberg and von Davier's (2020) idea of identifying unexpectedly high or low subscores by comparing observed subscores against a discrete distribution of subscores conditional on the overall proficiency variable in a unidimensional IRT model, but the computations and approach are different.

Our context is a psychiatric assessment tool—the Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001). PDSQ is a widely used self-report instrument. In particular, it is used in the well-known Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, a federally funded large-scale study comparing depression treatments. The instrument consists of 139 dichotomous items that cover 15 most prevalent DSM-IV (American Psychiatric Association, 1994) Axis I disorders. Using STAR*D data, which we also use here, Gibbons et al. (2009) showed that a bifactor model, which includes a general psychiatric distress dimension and 15 domain-specific latent dimensions, provides a plausible theoretical and statistical structure for the instrument.

In our preliminary analysis, three symptom domains—alcohol abuse dependence (ALC), drug abuse dependence (DRUG), and Psychosis (PSYCH)—are excluded. The exclusion of the first two is based on the empirical observation that the substance abuse domains were rather distinct from

TABLE 8.
Two-way lookup table that translates observed subscore combinations to composite scores

Listening score	Reading score																
	0	1	2	3	4	...	9	10	11	12	13	...	19	20	21	22	23
0	−3.96	−3.81	−3.65	−3.50	−3.35	...	−2.71	−2.59	−2.46	−2.35	−2.24	...	−1.75	−1.69	−1.64	−1.60	−1.56
1	−3.77	−3.62	−3.47	−3.33	−3.19	...	−2.59	−2.48	−2.36	−2.26	−2.15	...	−1.68	−1.62	−1.57	−1.53	−1.49
2	−3.58	−3.44	−3.30	−3.17	−3.04	...	−2.48	−2.37	−2.27	−2.17	−2.07	...	−1.61	−1.55	−1.50	−1.45	−1.42
3	−3.40	−3.27	−3.14	−3.02	−2.90	...	−2.38	−2.28	−2.18	−2.08	−1.99	...	−1.54	−1.48	−1.43	−1.39	−1.35
4	−3.24	−3.12	−3.00	−2.88	−2.78	...	−2.29	−2.19	−2.10	−2.01	−1.92	...	−1.47	−1.41	−1.36	−1.32	−1.28
5	−3.10	−2.98	−2.87	−2.76	−2.66	...	−2.20	−2.11	−2.02	−1.93	−1.85	...	−1.41	−1.35	−1.30	−1.26	−1.22
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10	−2.52	−2.43	−2.34	−2.27	−2.19	...	−1.84	−1.76	−1.69	−1.61	−1.54	...	−1.10	−1.05	−1.00	−.95	−.91
11	−2.43	−2.34	−2.26	−2.19	−2.12	...	−1.77	−1.70	−1.62	−1.55	−1.48	...	−1.05	−.99	−.94	−.90	−.86
12	−2.34	−2.26	−2.18	−2.11	−2.04	...	−1.70	−1.63	−1.56	−1.49	−1.42	...	−1.00	−.94	−.89	−.85	−.81
13	−2.25	−2.17	−2.10	−2.03	−1.96	...	−1.64	−1.57	−1.50	−1.44	−1.37	...	−.95	−.89	−.84	−.79	−.75
14	−2.17	−2.09	−2.02	−1.96	−1.89	...	−1.59	−1.52	−1.46	−1.39	−1.33	...	−.88	−.82	−.76	−.72	−.67
15	−2.09	−2.02	−1.96	−1.90	−1.84	...	−1.54	−1.47	−1.41	−1.34	−1.27	...	−.80	−.74	−.68	−.63	−.59
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
21	−1.76	−1.70	−1.64	−1.58	−1.52	...	−1.23	−1.17	−1.10	−1.03	−.95	...	−.34	−.25	−.17	−.10	−.04
22	−1.70	−1.64	−1.58	−1.53	−1.47	...	−1.18	−1.11	−1.04	−.96	−.87	...	−.20	−.10	−.01	.07	.14
23	−1.65	−1.59	−1.53	−1.47	−1.41	...	−1.11	−1.04	−.97	−.89	−.79	...	−.04	.07	.17	.26	.34
24	−1.59	−1.53	−1.47	−1.41	−1.35	...	−1.04	−.97	−.89	−.81	−.71	...	.14	.26	.38	.48	.57
25	−1.53	−1.46	−1.40	−1.34	−1.28	...	−.97	−.90	−.82	−.73	−.62	...	.34	.48	.61	.72	.83
26	−1.46	−1.40	−1.33	−1.27	−1.22	...	−.90	−.83	−.74	−.64	−.53	...	.57	.72	.87	1.00	1.12
27	−1.40	−1.33	−1.27	−1.21	−1.16	...	−.83	−.75	−.67	−.56	−.43	...	.82	.99	1.15	1.30	1.43

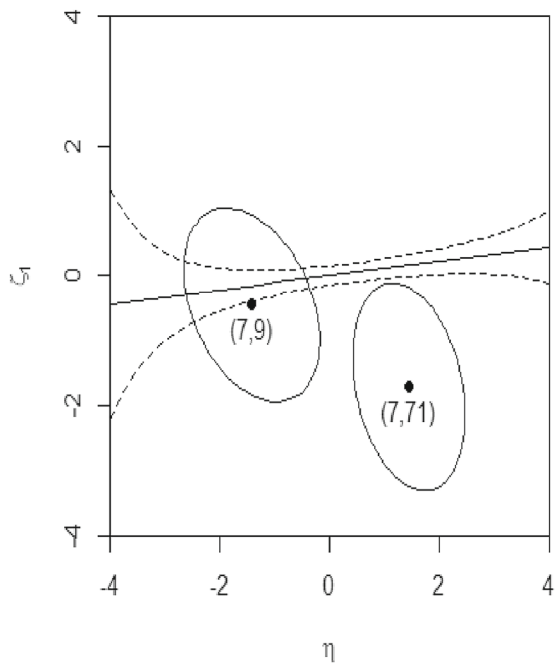

FIGURE 3.
95% prediction interval of η-ξ_1 regression

the other domains, as judged from the item slopes. The Psychosis domain is excluded because the STAR*D participants are screened positive for non-psychotic major depressive disorder. Two more items in the major depressive disorder (MDD) domain were further excluded due to their ill fit. Therefore, the MDD domain is measured by 19 dichotomous items, while the rest score on the other 11 domains can range from 0 to 100. Item parameters are calibrated based on a sample of 3999 participants and are assumed to be fixed in the subsequent analysis. The illustrative task is to identify combinations of observed scores on the MDD domain (i.e., s_1) versus the rest (i.e., $s_{(1)}$) that signal the reporting of the MDD subscore would add information to the overall score.

We note that the summaries of the posterior distribution (i.e., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) along with the probability associated with each observed score combination, obtainable from the Lord–Wingersky algorithm 2.5, could be utilized to capture the statistical relationship between ξ_n and η and thereafter facilitate subscore reporting. In the simplest instantiation, we regress the estimate of ξ_1 of each score combination $(s_1, s_{(1)})$ on the η estimate, weighted by the corresponding marginal probability, $p\left(s_1, s_{(1)}\right)$. A 95% prediction interval from this weighted least squares regression can be calculated (Fig. 3) with the regression parameter estimates and serves as the basis to evaluate the bivariate normal approximated posterior of each $(s_1, s_{(1)})$. For each score combination, the proportion of the posterior density volume that falls within the prediction interval is computed, akin to a p-value. The smaller the proportion, the more necessity there is to report the ξ_1 estimate associated with the score combination. For example, as in Fig. 3, the MDD subscore associated with (7, 71) should be reported, while for another combination, (7, 9), it may not be necessary. Table 9 shows proportions of posterior volumes that fall in the prediction interval, with darker cells indicating lower proportions.

TABLE 9.
Proportions of posterior volume that falls in prediction interval

s_1	$s_{(1)}$																					
	0	1	2	…	23	24	25	…	38	39	40	…	51	52	53	…	65	66	67	…	99	100
0	.19	.12	.07	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00
1	.29	.20	.14	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00
2	.38	.28	.21	…	.01	.01	.01	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00
3	.47	.37	.29	…	.02	.01	.01	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00
4	.53	.44	.37	…	.03	.03	.03	…	.01	.01	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00
5	.56	.49	.42	…	.06	.05	.05	…	.01	.01	.01	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00
6	.56	.50	.45	…	.09	.08	.08	…	.02	.02	.02	…	.01	.01	.01	…	.00	.00	.00	…	.00	.00
7	.50	.47	.44	…	.13	.12	.11	…	.04	.04	.03	…	.02	.02	.02	…	.01	.01	.01	…	.00	.00
8	.40	.40	.39	…	.17	.16	.15	…	.06	.06	.06	…	.03	.03	.03	…	.02	.02	.02	…	.00	.00
9	.26	.28	.30	…	.21	.20	.19	…	.09	.09	.09	…	.06	.06	.05	…	.03	.03	.03	…	.00	.00
10	.13	.16	.18	…	.23	.22	.21	…	.14	.13	.13	…	.10	.09	.09	…	.06	.06	.06	…	.00	.00
11	.05	.06	.08	…	.22	.22	.21	…	.18	.18	.17	…	.15	.15	.15	…	.11	.10	.09	…	.01	.01
12	.01	.02	.03	…	.17	.17	.18	…	.20	.21	.21	…	.22	.22	.22	…	.17	.17	.16	…	.02	.02
13	.00	.01	.01	…	.10	.10	.11	…	.18	.19	.20	…	.27	.28	.28	…	.26	.25	.24	…	.04	.05
14	.00	.00	.00	…	.04	.04	.05	…	.12	.12	.13	…	.25	.26	.27	…	.32	.31	.30	…	.13	.13
15	.00	.00	.00	…	.01	.02	.02	…	.06	.06	.07	…	.16	.17	.18	…	.26	.26	.26	…	.33	.35
16	.00	.00	.00	…	.01	.01	.01	…	.03	.03	.03	…	.08	.08	.09	…	.15	.15	.16	…	.58	.61
17	.00	.00	.00	…	.00	.00	.00	…	.01	.01	.01	…	.04	.04	.04	…	.07	.07	.08	…	.64	.69
18	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.01	…	.01	.01	.02	…	.03	.03	.03	…	.48	.54
19	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.00	.00	.00	…	.01	.01	.01	…	.31	.37

4.3. Detecting Aberrant Score Combination Pattern

As mentioned in Sect. 4.1, the probability of each observed score combination is a by-product of the Lord–Wingersky algorithm 2.5. When arranged in a contingency table, the probability of observed subscore combination $(s_n, s_{(n)})$ can be used to detect aberrant score combinations through the construction of posterior high-density region (HDR; Novick & Jackson, 1974). A low probability indicates that the co-occurrence of corresponding summed scores is rare. Depending on context, this approach can be useful for diagnosis of lack of person fit or for forensic data analysis in test security.

We illustrate this application of Lord–Wingersky algorithm 2.5 with the Quality of Life (QoL) Scale for the Chronically Mentally Ill (Lehman, 1988). Many previous studies indicate a bifactor model fits the 35-item QoL scale extremely well (e.g., Gibbons et al., 2007; Cai & Hansen, 2013). Beyond an overall quality of life item, there are 7 subscales (*Family*, *Finance*, *Health*, *Leisure*, *Living*, *Safety*, and *Social*), each of which includes 4 to 6 items. The dataset used here includes responses from 586 patients. To aid presentation, the original 7-category rating scale items are recoded to have two categories (i.e., 0, 1, and 2 in the original scale are recoded as 0; 3, 4, 5, and 6 as 2). Here, we construct the high-density region (HDR) of combinations of the score on *Health* (s_1) and the rest score $(s_{(1)})$ using the Lord–Wingersky algorithm 2.5. s_1 ranges from 0 to 6, and $s_{(1)}$ ranges from 0 to 29.

To construct a HDR of level α, we first stack the $p(s_1, s_{(1)})$ of each score combination into a single column, sort all the probabilities from the largest to the smallest, and then compute the cumulative distribution of these probabilities. Observed score combinations that contribute to the first $100\alpha\%$ of the cumulative distribution are identified as the $100\alpha\%$ HDR.

Figure 4 shows the HDR for the illustrative task. The unshaded cells represent the 95% HDR. The light gray cells together with the unshaded cells represent the 99% HDR. The dark gray cells represent observed subscore patterns that rarely occur. For example, the score combinations (0, 29) and (6, 0) rarely occur. Individuals with such score combinations deserve further attention.

FIGURE 4.
High-density region (*Health* versus the rest)

5. Discussion

The original Lord–Wingersky (1984) algorithm was developed for binary items under unidimensional IRT models. Then the algorithm was expanded to polytomous unidimensional IRT models (Hanson, 1994; Thissen et al., 1995; von Davier & Rost, 1995). The Lord–Wingersky algorithm version 2.0 (Cai, 2015) was proposed to computed likelihoods associated with overall summed scores in the context of hierarchical item factor models. In the present article, we proposed the Lord–Wingersky algorithm 2.5 as an extension of the Cai's (2015) Lord–Wingersky algorithm 2.0. The algorithm yields the characterization of the bivariate posterior associated with observed score combinations from the mutually exclusive clusters of items in the model. The algorithm uses more observed information than the Lord–Wingersky Algorithm 2.0 (observed score combinations instead of one overall summed score). Thus it can provide additional information that is useful in practice (summed score likelihoods for all latent dimension instead of the likelihood for the primary latent dimension only). The Lord–Wingersky algorithm 2.5 also remains computationally efficient due to the continued use of dimension reduction. With the Lord–Wingersky algorithm 2.5, likelihoods of observed score combinations under several IRT models, including the two-tier model, the bifactor model and the standard MIRT model, can be computed directly under one algorithm.

The bivariate normal approximation (summarized by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) to the posterior associated with each observed score combination, as one of the outputs of the Lord–Wingersky algorithm 2.5, is a reasonable alternative to the actual (intractable) posterior distribution and can serve multiple purposes in educational and psychological measurement. The marginal probability of each observed score combination, which comes as a by-product of the proposed algorithm, is also useful in practice. We use three empirical applications to illustrate the range of possible use of this new algorithm—(a) translating observed score combinations to aid growth interpretations in educational measurement, (b) facilitating subscore reporting in psychiatric assessment, and (c) detecting aberrant observed subscore combinations in health-related outcome research.

While applying the proposed algorithm, we assume the IRT model is correct. It is also assumed that item parameters are known and fixed, since in practice the parameter calibration stage and scoring stage are often conducted sequentially. To take into account the uncertainty around item parameters (i.e., standard errors in the calibration stage), we suggest using multiple imputation (MI; Rubin, 1987)-based approach (e.g., Yang et al., 2012).

Hierarchical item factor models, especially the bifactor model, saw increasing use in psychological and educational assessment. Recent development in computational algorithms for estimating multidimensional IRT models (Cai, 2010a; Edwards, 2010) and software, e.g., flexMIRT® (Cai, 2017), has brought the usage of MIRT models within reach for routine data analysis. We posit that providing scores that are based on observed statistics (e.g., summed scores, observed subscale scores) will continue to be desired and useful in practice, and the current study is a further contribution to the IRT scoring literature.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Author.
Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society-Series B*, *34*, 42–54.
Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.
Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57.
Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.
Cai, L. (2015). Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, *80*(2), 535–559.
Cai, L. (2017). *flexMIRT® : Flexible multilevel multidimensional item analysis and test scoring.* (version 3.51) [Computer software]. Vector Psychometric Group.
Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276.
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*(1), 19–37.
Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.
English Language Proficiency Assessment for the 21st Century. (2017). *Item analysis and calibration*. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Feinberg, R. A., & von Davier, M. (2020). Conditional subscore reporting using iterated discrete convolutions. *Journal of Educational and Behavioral Statistics*, *45*(5), 515–533.
Feinberg, R. A., & Wainer, H. (2014). A simple equation to predict a subscore's value. *Educational Measurement: Issues and Practice*, *33*(3), 55–56.
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*(1), 4–19.
Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.
Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of psychiatric research*, *43*(4), 401–410.
Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement*, *40*, 327–385.
Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229.
Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 79–95.
Hansen, M., & Monroe, S. (2018). Linking not-quite-vertical scales through multidimensional item response theory. *Measurement: Interdisciplinary Research and Perspectives*, *16*(3), 155–167.
Hanson B. A. (1994). *Extension of Lord–Wingersky algorithm to computing test score distributions for polytomous items*. Unpublished manuscript. Retrieved Jan 1, 2016, from? from http://www.b-a-h.com/papers/note9401.pdf
Kim, S. (2013). Generalization of the Lord–Wingersky algorithm to computing the distribution of summed test scores based on real-number item scores. *Journal of Educational Measurement*, *50*(4), 381–389.
Lehman, A. F. (1988). A quality of life interview for the chronically mentally ill. *Evaluation and Program Planning*, *11*(1), 51–62.
Li, Z., & Cai, L. (2018). Summed score likelihood-based indices for testing latent variable distribution fit in item response theory. *Educational and Psychological Measurement*, *78*(5), 857–886.
Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score" equatings". *Applied Psychological Measurement*, *8*(4), 453–461.
Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. McGraw-Hill.
Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*(3), 354.
Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.
Reckase, M. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences)*. Springer.
Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696.
Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor modelling and the evaluation of scale scores. In P. Irwing, T. Booth & D. J. Hughes (Eds.) *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 675–707). John Wiley & Sons Ltd.
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*, *16*, 19–31.
Rijmen, F. (2009). Efficient full information maximum likelihood estimation for multidimensional IRT models. *ETS Research Report Series*, *2009*(1), i–31.
Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiplechoice and constructed-response items-scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), *Test*

scoring (pp. 253–292). Lawrence Erlbaum.
Rubin, D. B. (1987). *Multiple imputations for nonresponse in surveys*. Wiley.
Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*, 21–28.
Stucky, B. D. (2009). *Item response theory for weighted summed scores*, Doctoral dissertation, The University of North Carolina at Chapel Hill.
Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39–49.
Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL^{TM} asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, *20*, 1497–1505.
Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Routledge.
von Davier, M. & Rost, J. (1995) Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.) *Rasch models: Foundations, recent developments and applications* (pp. 371–379). Springer.
Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.
Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290.
Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement*, *19*(3), 231–240.
Zimmerman, M., & Mattia, J. I. (2001). A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire. *Archives of General Psychiatry*, *58*(8), 787–794.

Manuscript Received: 19 OCT 2020
Final Version Received: 18 JUN 2021
Accepted: 25 JUN 2021
Published Online Date: 27 JUL 2021