Research Report

How Can Multivariate Item Response Theory Be Used in Reporting of Subscores?

Shelby J. Haberman

Sandip Sinharay

March 2010

ETS RR-10-09

How Can Multivariate Item Response Theory Be Used in Reporting of Susbcores?

Shelby J. Haberman and Sandip Sinharay

ETS, Princeton, New Jersey

March 2010

Technical Review Editor: Dan Eignor

Technical Reviewers: Frank Rijmen and Hongwen Guo

Abstract

Recently, there has been increasing interest in reporting diagnostic scores. This paper examines reporting of subscores using multidimensional item response theory (MIRT) models. An MIRT model is fitted using a stabilized Newton-Raphson algorithm (Haberman, 1974, 1988) with adaptive Gauss-Hermite quadrature (Haberman, von Davier, & Lee, 2008). A new statistical approach is proposed to assess when subscores using the MIRT model have any added value over (a) the total score or (b) subscores based on classical test theory (Haberman, 2008; Haberman, Sinharay, & Puhan, 2006). The MIRT-based methods are applied to several operational data sets. The results show that the subscores based on MIRT are slightly more accurate than subscore estimates derived by classical test theory.

Key words: 2PL model, mean-squared error, augmented subscore

Acknowledgments

The authors thank Frank Rijmen, Hongwen Guo, and Dan Eignor for useful advice, and Amanda McBride and Kim Fryer for editorial help.

There is an increasing interest in subscores because of their potential diagnostic value. Failing candidates want to know their strengths and weaknesses in different content areas to plan for future remedial work. States and academic institutions such as colleges and universities often want a profile of performance for their graduates to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004).

Multidimensional item response theory (MIRT) models can be employed to report subscores. Several papers have suggested this approach, although current approaches have been somewhat problematic in terms of practical application to testing programs with limited time for analysis. For instance, de la Torre and Patz (2005) applied an MIRT model to data from tests that measure multiple correlated abilities. This method can be used to estimate subscores, although the subscores, which are components of the ability vector in the MIRT model, are in the scale of the ability parameters rather than in the scale of the raw scores. This approach provided results very similar to those based on augmentation of raw subscores (Wainer et al., 2001). Yao and Boughton (2007) also examined subscore reporting based on an MIRT model and the Markov-chain Monte-Carlo (MCMC) algorithm. However, the MCMC algorithm employed in de la Torre and Patz (2005) or Yao and Boughton (2007) is more computationally intensive than is currently practical given the time constraints of many testing programs. In addition, determination of convergence of an MCMC algorithm is not straightforward for a typical psychometrician working for a testing company. Researchers have also compared different approaches, including the MIRT-based methods, for reporting subscores. For example, Dwyer, Boughton, Yao, Steffen, and Lewis (2006) compared four methods: raw subscores, the objective performance index (OPI) described in Yen (1987), Wainer augmentation, and MIRT-based subscores. On the whole, they found that the MIRT-based methods and augmentation methods provided the best estimates of subscores.

This paper fits the MIRT model using a stabilized Newton-Raphson algorithm (Haberman, 1974, 1988) with adaptive Gauss-Hermite quadrature (Haberman, von Davier, & Lee, 2008). In typical applications, this algorithm is far faster than the MCMC algorithm, so that methods used in this paper can be considered in operational testing. In addition, a new statistical approach is proposed to assess when subscores obtained using MIRT have any added value over (a) the total score and (b) subscores based on classical test theory. This work extends to MIRT models the research of Haberman (2008) and Haberman, Sinharay, and Puhan (2006), who suggested methods based on classical test theory (CTT) to examine whether subscores provide any added value over total scores.

Section 1 of this report provides a brief overview of the CTT-based methods of Haberman (2008) and Haberman et al. (2006). Section 2 introduces the MIRT model under study, suggests how to compute the subscores based on MIRT, and suggests how to assess when subscores using MIRT have any added value over the total score and over subscores based on classical test theory. Section 3 illustrates application of the methods to several data sets. Section 4 provides conclusions based on the empirical results observed.

Discussion in this report is confined to right-scored tests in which subscores of interest do not share common items. Adaptation to tests with polytomous items is straightforward. Treatment of subscores with overlapping items is somewhat more complicated. The authors plan to report on this case in a future publication.

1 Methods From Classical Test Theory

This section describes the approach of Haberman (2008) and Haberman et al. (2006) to determine whether and how to report subscores. Consider a test with $q \geq 2$ right-scored items. A sample of $n \geq 2$ examinees is used in analysis of the data. For examinee i, $1 \leq i \leq n$, and for item j, $1 \leq j \leq q$, X_{ij} is 1 if the response to item j is correct, and X_{ij} is 0 otherwise. The q-dimensional vectors $\mathbf{X}_i$ with coordinates X_{ij}, $1 \leq j \leq q$, are independent and identically distributed for examinees i from 1 to n, and the set of possible values of $\mathbf{X}_i$ is denoted by Γ. The items test $r \geq 2$ skills numbered from 1 to r. To each item j, $1 \leq j \leq q$, corresponds a single skill $v(j)$, $1 \leq v(j) \leq r$. It is assumed that each skill corresponds to some item. Thus, if $J(k)$ denotes the set of items j with skill $v(j) = k$, then $J(k)$ is nonempty for $1 \leq k \leq r$.

In a CTT-based analysis, examinee i has total raw score

$$S_i = \sum_{j=1}^{q} X_{ij}$$

and raw subscore

$$S_{ik} = \sum_{j \in J(k)} X_{ij},$$

which corresponds to skill k. The true score corresponding to S_i is the true total raw score T_i, and the true score corresponding to S_{ik} is the true raw subscore T_{ik}. Proposed subscores are judged by how well they approximate the true subscores T_{ik}. The following subscores are considered for examinee i and skill k:

- The linear combination $U_{iks} = \alpha_{ks} + \beta_{ks} S_{ik}$ based on the raw subscore S_{ik}, which yields the minimum (denoted as τ^2_{ks}) of the mean-squared error $E([T_{ik} - U_{iks}]^2)$.
- The linear combination $U_{ikx} = \alpha_{kx} + \beta_{kx} S_i$ based on the raw total score S_i, which yields the minimum (τ^2_{kx}) of the mean-squared error $E([T_{ik} - U_{ikx}]^2)$.
- The linear combination $U_{ikc} = \alpha_{kc} + \beta_{k1c} S_i + \beta_{k2c} S_{ik}$ based on the raw subscore S_{ik} and raw total score S_i, which yields the minimum (τ^2_{kc}) of the mean-squared error $E([T_{ik} - U_{ikc}]^2)$.

The subscore U_{ikc} is an example of an augmented subscore (Wainer et al., 2001). We will often refer to the procedure by which U_{ikc} is obtained as the Haberman augmentation. It is also possible to consider an augmented subscore $U_{ika} = \alpha_{ka} + \sum_{k'=1}^{r} \beta_{kk'a} S_{ik'}$ based on all the raw subscores (Wainer et al., 2001), which yields the minimum τ^2_{ka} of the mean-squared error $E([T_{ik} - U_{ika}]^2)$. Because this augmentation typically provides results that are very similar to those of Haberman augmentation, we do not provide any results for U_{ika} in this paper.

To compare the possible subscores, proportional reduction in mean-squared error (PRMSE) is employed. Let τ^2_{k0} be the variance of the true raw subscore T_{ik}, so that τ^2_{k0} is the minimum of $E([T_{ik} - U_{ik0}]^2)$ for the constant approximation $U_{ik0} = a_{k0}$. Then τ^2_{ks}, τ^2_{kx}, τ^2_{kc}, and τ^2_{ka} cannot exceed τ^2_{k0}. The proportional reductions of mean-squared error for the subscores under study are

$$\mathrm{PRMSE}_{ks} = 1 - \tau^2_{ks}/\tau^2_{k0},$$

$$\mathrm{PRMSE}_{kx} = 1 - \tau^2_{kx}/\tau^2_{k0},$$

$$\mathrm{PRMSE}_{kc} = 1 - \tau^2_{kc}/\tau^2_{k0},$$

and

$$\mathrm{PRMSE}_{ka} = 1 - \tau^2_{ka}/\tau^2_{k0}.$$

The reliability coefficient of S_{ik} is PRMSE_{ks}. Each PRMSE is between 0 and 1. Because reduced mean-squared error is desired, it is clearly best to have a PRMSE close to 1. It is always the case that $\mathrm{PRMSE}_{ks} \leq \mathrm{PRMSE}_{kc}$, $\mathrm{PRMSE}_{kx} \leq \mathrm{PRMSE}_{kc}$, and $\mathrm{PRMSE}_{kc} \leq \mathrm{PRMSE}_{ka}$.

Consideration of the competing interests of simplicity and accuracy suggests the following strategy (Haberman, 2008; Haberman et al., 2006) for skill k:

- If PRMSE_{ks} is less than PRMSE_{kx}, declare that the subscore *does not provide added value over the total score.*

- Use U_{kc} only if PRMSE_{kc} is substantially larger than the maximum of PRMSE_{ks} and PRMSE_{kx}.

The first recommendation reflects the fact that the observed total score will provide more accurate diagnostic information than the observed subscore if PRMSE_{ks} is less than PRMSE_{kx}. Sinharay, Haberman, and Puhan (2007) discussed the strategy in terms of reasonableness and in terms of compliance with professional standards. The second recommendation involves the slight increase in computation when U_{kc} is employed and the challenges in explaining score augmentation to clients. In practice, use of U_{iks} is most attractive if the raw subscore S_{ik} has high reliability and if the correlations of the true raw subscores are not very high (Haberman, 2008; Haberman et al., 2006).

Haberman (2008) discussed the estimation from sample data of the proposed subscores, the regression coefficients, the mean-squared errors, and PRMSE coefficients. The straightforward computations depend only on the sample moments and correlations among the subscores and their reliabilities. For large samples, the decrease in PRMSE due to estimation is negligible.

2 The Two-Parameter Logistic (2PL) MIRT Model

The two-parameter logistic (2PL) MIRT model employed in this report is a simple-structure model described in Haberman et al. (2008). The basic 2PL MIRT model under study assumes that an r-dimensional random ability vector $\boldsymbol{\theta}_i$ with coordinates θ_{ik}, $1 \le k \le r$ is associated with each examinee i. The pairs $(\mathbf{X}_i, \boldsymbol{\theta}_i)$, $1 \le i \le n$ are independent and identically distributed, and, for each examinee i, the response variables X_{ij}, $1 \le j \le q$, are conditionally independent given $\boldsymbol{\theta}_i$. Let

$$P(h; y) = \exp(hy)/[1 + \exp(y)]$$

for h and y real.

To each item j, $1 \le j \le q$, the conditional probability that $X_{ij} = h$ given $\boldsymbol{\theta}_i = \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is an r-dimensional vector of real numbers, is $P(h; a_j\omega_{\upsilon(j)} - \gamma_j)$ for an unknown item discrimination a_j and an unknown real parameter γ_j. Provided that the discrimination a_j is positive, the item difficulty for item j is then $\gamma_j/a_j = b_j$.The conditional probability that $\mathbf{X}_i = \mathbf{x}$ given that $\boldsymbol{\theta}_i$ is equal to the r-dimensional vector $\boldsymbol{\omega}$ is then

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{q} P(h; a_j \omega_{\upsilon(j)} - \gamma_j). \tag{1}$$

If a_j is constant for j in $J(k)$, $1 \le k \le r$, then one has a multidimensional one-parameter logistic (1PL) model.

In this report, the assumption is made that $\boldsymbol{\theta}_i$ has a multivariate normal distribution $N(\mathbf{0}, \mathbf{D})$. Here $\mathbf{0}$ is the r-dimensional vector with all coordinates 0, and $\mathbf{D}$ is an r-by-r positive-definite symmetric matrix with elements $d_{kk'}$, $1 \le k \le r$, $1 \le k' \le r$, such that each diagonal element d_{kk} is equal to 1, and each off-diagonal element $d_{kk'}$, $k \ne k'$, is the unknown correlation of θ_{ik} and $\theta_{ik'}$. The assumption that the mean of $\boldsymbol{\theta}_i$ is $\mathbf{0}$ and the variance d_{kk} of each θ_{ik} is 1 is imposed to permit identification of the item parameters a_j and b_j for each item j from 1 to q. Alternative analysis is possible in which other distributions of $\boldsymbol{\theta}_i$ are considered (Haberman et al., 2008).

The model parameters a_j and γ_j, $1 \le j \le q$, and $d_{kk'}$, $1 \le k < k' \le r$, may be estimated by maximum-likelihood by means of a version of the stabilized Newton-Raphson algorithm (Haberman, 1988) described in Haberman et al. (2008). Because calculations employ adaptive multivariate Gauss-Hermite integration, computational time is not excessive (Schilling & Bock, 2005).

The maximum-likelihood estimates $\hat{a}_j$ of a_j, $\hat{\gamma}_j$ of γ_j, and $\hat{\mathbf{D}}$ of $\mathbf{D}$ continue to estimate meaningful parameters even if the model does not hold because a_j, γ_j, and $\mathbf{D}$ can be selected to minimize the expected log penalty function $E(-\log p(\mathbf{X}_i))$ for $p(\mathbf{x})$, $\mathbf{x}$ in Γ, the expected value of $p(\mathbf{x}|\boldsymbol{\theta}_i)$ (Gilula & Haberman, 1994, 2001; Haberman, 2007). In this fashion, a_j, γ_j, and $\mathbf{D}$ can be regarded as the parameters that result in the best correspondence between the model and the actual probability distribution of the response vector $\mathbf{X}$. If the model holds, then the optimal a_j and γ_j are the model parameters in (1), and the optimal $\mathbf{D}$ is the covariance matrix of $\boldsymbol{\theta}_i$.

Given the general definition of the model parameters in terms of expected log penalty, the ability parameter $\boldsymbol{\theta}_i$ can be defined and approximated even if the underlying model is not accurate (Haberman, 2007). To do so, let $\boldsymbol{\theta}_i$ be defined as a random vector such that the conditional distribution of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i = \mathbf{x}$ is the same as the conditional distribution of a random vector $\boldsymbol{\theta}_i^*$ given the random vector $\mathbf{X}_i^*$ with values in Γ, where $\boldsymbol{\theta}_i^*$ has a multivariate normal distribution with zero mean and covariance matrix $\mathbf{D}$ and the conditional probability that $\mathbf{X}_i^* = \mathbf{x}$ in Γ given $\boldsymbol{\theta}_i^* = \boldsymbol{\omega}$ is $p(\mathbf{x}|\boldsymbol{\omega})$. Let π denote the density function of a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$, so that $p(\mathbf{x})$ is the integral $\int p(\mathbf{x}|\boldsymbol{\omega})\pi(\boldsymbol{\omega})d\boldsymbol{\omega}$. By Bayes's theorem, the conditional density of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i = \mathbf{x}$ has value

$$f(\boldsymbol{\omega}|\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\omega})\pi(\boldsymbol{\omega})/p(\mathbf{x})$$

at the r-dimensional vector $\boldsymbol{\omega}$. The unconditional density of $\boldsymbol{\theta}_i$ at $\boldsymbol{\omega}$ is then

$$f(\boldsymbol{\omega}) = E(f(\boldsymbol{\omega}|\mathbf{X}_i)).$$

If the model holds, then $f(\boldsymbol{\omega}) = \pi(\boldsymbol{\omega})$.

The expected a posteriori (EAP) mean $\tilde{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i$ (Bock & Aitkin, 1981) is the basis for the analysis of subscores by multivariate item response models. This mean is $\int \boldsymbol{\omega} f(\boldsymbol{\omega}|\mathbf{X}_i)d\boldsymbol{\omega}$. Clearly $\tilde{\boldsymbol{\theta}}_i$ has expectation $E(\boldsymbol{\theta}_i) = \int \boldsymbol{\omega} f(\boldsymbol{\omega})d\boldsymbol{\omega}$. The covariance matrix of $\boldsymbol{\theta}_i$ is

$$\mathrm{Cov}(\boldsymbol{\theta}_i) = \int [\boldsymbol{\omega} - E(\boldsymbol{\theta}_i)][\boldsymbol{\omega} - E(\boldsymbol{\theta}_i)]' f(\boldsymbol{\omega})d\boldsymbol{\omega},$$

where the prime indicates a transpose, while the approximation error $\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i$ has zero mean and covariance matrix

$$\mathrm{Cov}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) = E(\mathrm{Cov}(\boldsymbol{\theta}_i|\mathbf{X}_i)),$$

where

$$\mathrm{Cov}(\boldsymbol{\theta}_i|\mathbf{x}) = \int (\boldsymbol{\omega} - \tilde{\boldsymbol{\theta}}_i)(\boldsymbol{\omega} - \tilde{\boldsymbol{\theta}}_i)' f(\boldsymbol{\omega}|\mathbf{x})d\boldsymbol{\omega}.$$

For $1 \le k \le r$, let the coordinate vector $\boldsymbol{\delta}_k$ be the r-dimensional vector with coordinates

$$\delta_{k'k} = \begin{cases} 1, & k' = k, \\ 0, & k' \ne k, \end{cases}$$

for $1 \le k' \le k$. The kth coordinate θ_{ik} of $\boldsymbol{\theta}_i$ has variance $\tau^2_{k0\theta} = \boldsymbol{\delta}'_k \mathrm{Cov}(\boldsymbol{\theta}_i)\boldsymbol{\delta}_k$, and, for the kth coordinate $\tilde{\theta}_{ik}$ of $\tilde{\boldsymbol{\theta}}_i$, the mean-squared error $\tau^2_{k\theta}$ is

$$E([\theta_{ik} - \tilde{\theta}_{ik}]^2) = \boldsymbol{\delta}'_k E(\mathrm{Cov}(\boldsymbol{\theta}_i|\mathbf{X}_i))\boldsymbol{\delta}_k.$$

If the model holds, then $E(\boldsymbol{\theta}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\theta}_i) = \mathbf{D}$, so that $\tau^2_{k0\theta} = 1$.

For any nonzero fixed r-dimensional vector $\mathbf{c}$, the reliability of $\mathbf{c}'\tilde{\boldsymbol{\theta}}_i$ is then

$$\rho^2(\mathbf{c}) = 1 - \frac{\mathbf{c}' \mathrm{Cov}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\mathbf{c}}{\mathbf{c}' \mathrm{Cov}(\boldsymbol{\theta}_i)\mathbf{c}}. \tag{2}$$

The quantity $\mathbf{c}'\,\mathrm{Cov}(\tilde{\boldsymbol\theta}_i - \boldsymbol\theta_i)\mathbf{c}$ in (2) is both the variance of $\mathbf{c}'(\tilde{\boldsymbol\theta}_i - \boldsymbol\theta_i)$, where $\mathbf{c}'(\tilde{\boldsymbol\theta}_i - \boldsymbol\theta_i)$ can be considered as the error in approximation of $\mathbf{c}'\boldsymbol\theta_i$ by $\mathbf{c}'\tilde{\boldsymbol\theta}_i$, and the mean-squared error from approximation of $\mathbf{c}'\boldsymbol\theta_i$ by $\mathbf{c}'\tilde{\boldsymbol\theta}_i$. Similarly, $\mathbf{c}'\,\mathrm{Cov}(\boldsymbol\theta_i)\mathbf{c}$ in (2) is both the variance of $\mathbf{c}'\boldsymbol\theta_i$ and the minimum possible mean-squared error from approximation of $\mathbf{c}'\boldsymbol\theta_i$ by a constant. Thus, $\rho^2(\mathbf{c})$ has the form

$$\rho^2(\mathbf{c}) = 1 - \frac{\text{Error variance}}{\text{Total variance}},$$

which is the standard definition of reliability, and also has the form

$$\rho^2(\mathbf{c}) = \frac{\text{Reduction in MSE from approximation of } \mathbf{c}'\boldsymbol\theta_i \text{ by } \mathbf{c}'\tilde{\boldsymbol\theta}_i \text{ instead of by a constant}}{\text{MSE from approximation of } \mathbf{c}'\boldsymbol\theta_i \text{ by a constant}},$$

which is the usual form of a PRMSE (Haberman, 2008; Haberman et al., 2006).

It follows that the PRMSE for the kth coordinate $\tilde\theta_{ik}$ of $\tilde{\boldsymbol\theta}$ is

$$\mathrm{PRMSE}_{k\theta} = \rho^2(\boldsymbol\delta_k) = 1 - \tau^2_{k\theta}/\tau^2_{k0\theta}.$$

In practice, $\tilde{\boldsymbol\theta}_i$ must be approximated by

$$\hat{\boldsymbol\theta}_i = \int \boldsymbol\omega \hat f(\boldsymbol\omega|\mathbf{X}_i)d\boldsymbol\omega,$$

where

$$\hat f(\boldsymbol\omega|\mathbf{x}) = \hat p(\mathbf{x}|\boldsymbol\omega)\hat\pi(\boldsymbol\omega)/\hat p(\mathbf{x}),$$

$$\hat p(\mathbf{x}|\boldsymbol\theta) = \prod_{i=1}^{q} P(h; \hat a_j \omega_{v(j)} - \hat\gamma_j),$$

$\hat\pi$ is the density of a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\hat{\mathbf{D}}$, and

$$\hat p(\mathbf{x}) = \int \hat p(\mathbf{x}|\boldsymbol\omega)\hat\pi(\boldsymbol\omega)d\boldsymbol\omega.$$

For large samples, the reliability for $\hat{\boldsymbol\theta}_i$ is approximated by

$$\hat\rho^2(\mathbf{c}) = 1 - \frac{\mathbf{c}'\widehat{\mathrm{Cov}}(\tilde{\boldsymbol\theta} - \boldsymbol\theta)\mathbf{c}}{\mathbf{c}'\widehat{\mathrm{Cov}}(\boldsymbol\theta)\mathbf{c}},$$

where

$$\widehat{\mathrm{Cov}}(\tilde{\boldsymbol\theta} - \boldsymbol\theta) = n^{-1}\sum_{i=1}^{n}\widehat{\mathrm{Cov}}(\boldsymbol\theta_i|\mathbf{X}_i),$$

$$\widehat{\mathrm{Cov}}(\boldsymbol\theta_i|\mathbf{X}_i) = \int (\boldsymbol\omega - \hat{\boldsymbol\theta}_i)(\boldsymbol\omega - \hat{\boldsymbol\theta}_i])\hat f(\boldsymbol\omega|\mathbf{X}_i)d\boldsymbol\omega,$$

$$\widehat{\mathrm{Cov}}(\boldsymbol{\theta}) = \int (\boldsymbol{\omega} - \bar{\boldsymbol{\theta}})(\boldsymbol{\omega} - \bar{\boldsymbol{\theta}})'\hat{f}(\boldsymbol{\omega})d\boldsymbol{\omega},$$

$$\bar{\boldsymbol{\theta}} = \int \boldsymbol{\omega}\hat{f}(\boldsymbol{\omega})d\boldsymbol{\omega} = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i,$$

and

$$\hat{f}(\boldsymbol{\omega}) = n^{-1}\sum_{i=1}^{n}\hat{f}(\boldsymbol{\omega}|\mathbf{X}_i).$$

The estimated variance $\hat{\tau}^2_{k0\theta} = \boldsymbol{\delta}'_k\widehat{\mathrm{Cov}}(\boldsymbol{\theta})\boldsymbol{\delta}_k$, and, for the kth coordinate $\tilde{\theta}_{ik}$ of $\tilde{\boldsymbol{\theta}}_i$, the estimated mean-squared error $\hat{\tau}^2_{k\theta}$ is $\boldsymbol{\delta}'_k\widehat{\mathrm{Cov}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\boldsymbol{\delta}_k$. The estimated PRMSE is $\widehat{\mathrm{PRMSE}}_{k\theta} = 1 - \hat{\tau}^2_{k\theta}/\hat{\tau}^2_{k0\theta}$.

3 Applications

We analyzed data containing examinee responses from five tests used for educational certification. All of these tests report subscores operationally, and our goal here was to determine the best possible way to report subscores for these tests. To fit the MIRT model given by (1), we used a FORTRAN 95 program written by the lead author; the program uses a variation of the stabilized Newton-Raphson algorithm (Haberman, 1988) described in Haberman et al. (2008). Required quadratures are performed by adaptive multivariate Gauss-Hermite integration (Schilling & Bock, 2005).

3.1 Data Sets

The tests considered here contained only multiple-choice (MC) items and represented a broad range of content and skill areas such as elementary education, reading, writing, mathematics, and foreign languages. Results from this study may provide useful information for other tests of similar format and content. For confidentiality reasons, hypothetical names (e.g., Test A–E) are used for the tests. The number of items in each subscore for the five tests are presented in Tables 1–5. A brief description of the tests and the operationally reported subscores for each test is presented below. All of these data sets were considered in Puhan, Sinharay, Haberman, and Larkin (2008).

Test A is designed for prospective teachers of children in primary through upper elementary school grades. The 119 multiple-choice questions focus on four major subject areas: language arts/reading (30 items), mathematics (29 items), social studies (30 items), and science (30 items). The sample size (the number of examinees who took the form of Test A considered here) was 31,001, and the reliability of the total test score was 0.91.

Test B is designed for examinees who plan to teach in a special-education program at any grade level from preschool through grade 12. The 60 multiple-choice questions assess the examinee's knowledge of three major content areas: understanding exceptionalities (13 items), legal and societal issues (10 items), and delivery of services to students with disabilities (33 items). The sample size was 7,930, and the reliability of the total test score was 0.74.

Test C is designed to assess the knowledge and competencies necessary for a beginning or entry-year teacher of Spanish. This test consists of 116 MC questions organized into four broad categories: interpretive listening (31 items), structure of the language (35 items), interpretive reading (30 items), and cultural perspectives (20 items). The sample size was 2,154 and the reliability of the total test score was 0.94.

Test D is designed to assess the mathematical knowledge and competencies necessary for a beginning teacher of secondary school mathematics. It consists of 50 MC questions arragned into three broad categories, namely, mathematical concepts and reasoning (17 items), ability to integrate knowledge of different areas of mathematics (12 items), and the ability to develop mathematical models of real-life situations (21 items). The sample size was 6,818, and the reliability of the total test score was 0.82.

Test E is used to measure skills necessary for prospective and practicing paraprofessionals. It consists of 73 MC questions arranged into three broad categories: reading (25 items), mathematics (23 items), and writing (25 items). The sample size was 3,637, and the reliability of the total test score was 0.94.

3.2 Results

Tables 1 through 5 provide results for Tests A through E. Each of these tables shows the following:

- the number of items in the subscores,
- the estimated correlation between the raw subscores (simple and disattenuated),
- the estimated correlation $d_{kk'}$ between the components θ_{ik} and $\theta_{ik'}$ under the model, and
- the estimates of PRMSE$_{ks}$ (the subscore reliability), PRMSE$_{kx}$, PRMSE$_{kc}$, and PRMSE$_{k\theta}$.

These tables do not provide the names of the subscores (they are given earlier in the *Data Sets* subsection) and only denotes the subscores as Subscores 1, 2, ... Note that a comparison between

PRMSE$_{kc}$ and PRMSE$_{k\theta}$ will reveal wheter the MIRT approach provides subscores that, relative to their variability, are more accurate than those provided by the CTT approach.

Table 1
Results for Test A

	Subscores			
	1	2	3	4
Length	30	30	29	30
Correlation between the raw subscores	1.00	0.59	0.58	0.59
	0.78	1.00	0.53	0.60
	0.80	**0.68**	1.00	0.64
	0.84	**0.78**	**0.88**	1.00
Correlation between the components of $\boldsymbol{\theta}_i$	1.00			
	0.80	1.00		
	0.84	0.71	1.00	
	0.87	0.80	0.89	1.00
PRMSE$_{ks}$	0.71	0.83	0.73	0.71
PRMSE$_{kx}$	0.77	0.74	0.75	0.82
PRMSE$_{kc}$	0.82	0.86	0.82	0.84
PRMSE$_{k\theta}$	0.84	0.87	0.85	0.87

Note. In the correlation matrix between the raw subscores, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

The pattern of results is quite consistent. The MIRT subscores almost always yield a PRMSE at least as high as those provided by the augmented subscores. The differences are often quite small, but they are appreciable in a number of cases.

To investigate further the relationship between the MIRT subscores and the augmented subscores, Figures 1 and 2 provide, for each of the 4 subscores of Test C and for each of the 3 subscores of the Test D, (a) scatterplots of augmented subscores versus raw subscores (the panels in the top row), (b) the MIRT subscores versus the raw subscores (the panels in the middle row), and (c) the MIRT subscores versus the augmented subscores (the panels in the bottom row) for 1,000 randomly chosen examinees. Each panel also shows the correlation between the variables being plotted. Results were similar for the other tests and are not shown. While the correlations between the raw subscores and the augmented/MIRT subscores are between 0.86 and 0.97, the

Table 2

Results for Test B

	Subscores		
	1	2	3
Length	13	10	33
Correlation between the raw subscores	1.00	0.34	0.51
	0.96	1.00	0.41
	0.95	**0.99**	1.00
Correlation between the components of $\boldsymbol{\theta}_i$	1.00		
	0.96	1.00	
	0.96	0.94	1.00
PRMSE$_{ks}$	0.46	0.28	0.63
PRMSE$_{kx}$	0.71	0.73	0.73
PRMSE$_{kc}$	0.71	0.73	0.73
PRMSE$_{k\theta}$	0.74	0.71	0.75

Note. In the correlation matrix between the raw subscores, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

correlation between the MIRT subscores and the augmented subscores are very close to 1. In other words, there is a nearly perfect linear relationship between the MIRT subscores and the augmented subscores. Our finding of the similarity of the MIRT subscores and the augmented subscores supports the finding in de la Torre and Patz (2005) of the similarity of MCMC-based MIRT subscores and Wainer-augmented subscores. Figure 1 shows a curvilinear relationship between the raw subscores and the augmented/MIRT subscores. Figure 3, which shows histograms of the distributions of the raw subscores, augmented subscores, and MIRT subscores for Test C, shows substantial negative skewness in the distribution of subscores (due to several examinees obtaining maximum possible subscores);[1] this is the reason of the curvilinear relationship in Figure 1.

The results indicate that Haberman augmentation and the MIRT results strongly dominate the results for estimates that are based only on raw subscores. The augmented subscores and the MIRT-based subscores improve on the raw subscores and the total score with respect to PRMSE for Tests A, C, and E. Interestingly, for Test D, the augmented subscores do not improve on the total score with respect to PRMSE, but the MIRT-based subscores do. For Test B, neither the augmented subscores nor the MIRT-based subscores lead to any improvement over the total score.

Table 3
Results for Test C

	Subscores			
	1	2	3	4
Length	31	35	30	20
Correlation between the raw subscores	1.00	0.70	0.79	0.53
	0.85	1.00	0.73	0.55
	0.93	**0.87**	1.00	0.58
	0.70	**0.73**	**0.75**	1.00
Correlation between the components of $\boldsymbol{\theta}_i$	1.00			
	0.91	1.00		
	0.95	0.93	1.00	
	0.75	0.77	0.80	1.00
PRMSE_{ks}	0.84	0.83	0.86	0.68
PRMSE_{kx}	0.85	0.84	0.88	0.64
PRMSE_{kc}	0.89	0.88	0.91	0.77
$\text{PRMSE}_{k\theta}$	0.90	0.90	0.91	0.78

Note. In the correlation matrix between the raw subscores, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

The subscores are too unreliable for any diagnostic score reporting for this test.

4 Conclusions

The use of MIRT models to generate subscores is quite feasible, as evidenced by the examples. Given the similarity of results in terms of PRMSE to those from the CTT-based Haberman augmentation, client preferences may be a significant consideration. For clients preferring IRT models over CTT, this paper will provide a rational and practical approach to reporting subscores.

Computational burden for the MIRT analysis appears acceptable—the software program did not take more than a couple of hours to complete the calculations for any of the data sets we analyzed here. Several calculation details can be modified for much larger samples. The six quadrature points per dimension were somewhat higher than appears needed (Haberman et al., 2008). For example, for four dimensions, a reduction from six to three points per dimension reduces computational labor by a factor of about 16. In addition, it is often advisable to begin calculations with a few hundred or few thousand observations to establish good approximations of

Table 4
Results for Test D

	Subscores		
	1	2	3
Length	17	12	20
Correlation between the raw subscores	1.00	0.57	0.61
	0.95	1.00	0.58
	0.97	**0.94**	1.00
Correlation between the components of $\boldsymbol{\theta}_i$	1.00		
	0.92	1.00	
	0.97	0.93	1.00
PRMSE_{ks}	0.61	0.59	0.65
PRMSE_{kx}	0.81	0.78	0.81
PRMSE_{kc}	0.81	0.79	0.81
$\text{PRMSE}_{k\theta}$	0.83	0.81	0.84

Note. In the correlation matrix between the raw subscores, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

maximum-likelihood estimates. The approximations would then be used to complete computations with the full sample. Even with improved numerical techniques, the MIRT-based approach to computing subscores involves a much higher computational burden than is required for the CTT-based approach of Haberman (2008) and Haberman et al. (2006).

Use of the MIRT-based approach results in estimates that are more difficult to explain than are raw scores, although this issue can be alleviated by alternative scalings. For example, the conditional expectation of θ_{ik} given $\mathbf{X}_i$ could be replaced by the conditional expectations of $g_k(\theta_{ik})$ given $\mathbf{X}_i$, where, for real ω, $g_k(\omega)$ is the test characteristic curve

$$g_k(\omega) = \sum_{j \in J(k)} P(1; a_j\omega - \gamma_j)$$

corresponding to S_{ik}, so that $g_k(\omega)$ is the conditional expectation of S_{ik} given $\theta_{ik} = \omega$, and $g_k(\theta_{ik})$ is the true score corresponding to S_{ik} if the model is valid. See ? (?)habsinpsych) for further details on this issue.

MIRT-based estimates such as $\tilde{\theta}_{ik}$ are not on the same scale as the raw subscores S_{ik}. This affects comparisons of mean-squared or root mean-square errors but does not affect comparisons of

Table 5
Results for Test E

	Subscores		
	1	2	3
Length	25	23	25
Correlation between the raw subscores	1.00	0.76	0.79
	0.90	1.00	0.73
	0.91	**0.86**	1.00
Correlation between the components of $\boldsymbol{\theta}_i$	1.00		
	0.92	1.00	
	0.94	0.90	1.00
PRMSE_{ks}	0.87	0.84	0.85
PRMSE_{kx}	0.90	0.85	0.87
PRMSE_{kc}	0.91	0.89	0.90
$\text{PRMSE}_{k\theta}$	0.91	0.89	0.90

Note. In the correlation matrix between the raw subscores, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

PRMSE measures because any particular PRMSE is a dimensionless measure in which numerator and denominator are on the same scale.

Subscores must be reported on some established scale. A temptation exists to make this scale comparable to the scale for the total score or to the fraction of the scale that corresponds to the relative importance of the subscore, but these choices are not without difficulties given that subscores and total scores typically differ in reliability. In addition, if the subscore is worth reporting at all, then the subscore presumably does not measure the same construct as the total score. Further, appropriate methods of equating or linking must be considered when determining whether and how to report subscores. In typical cases, equating is feasible for the total score but not for subscores. For example, if an anchor test is used to equate the total test, only a few of the items will correspond to a particular subscore, so anchor test equating of the subscore is not feasible.

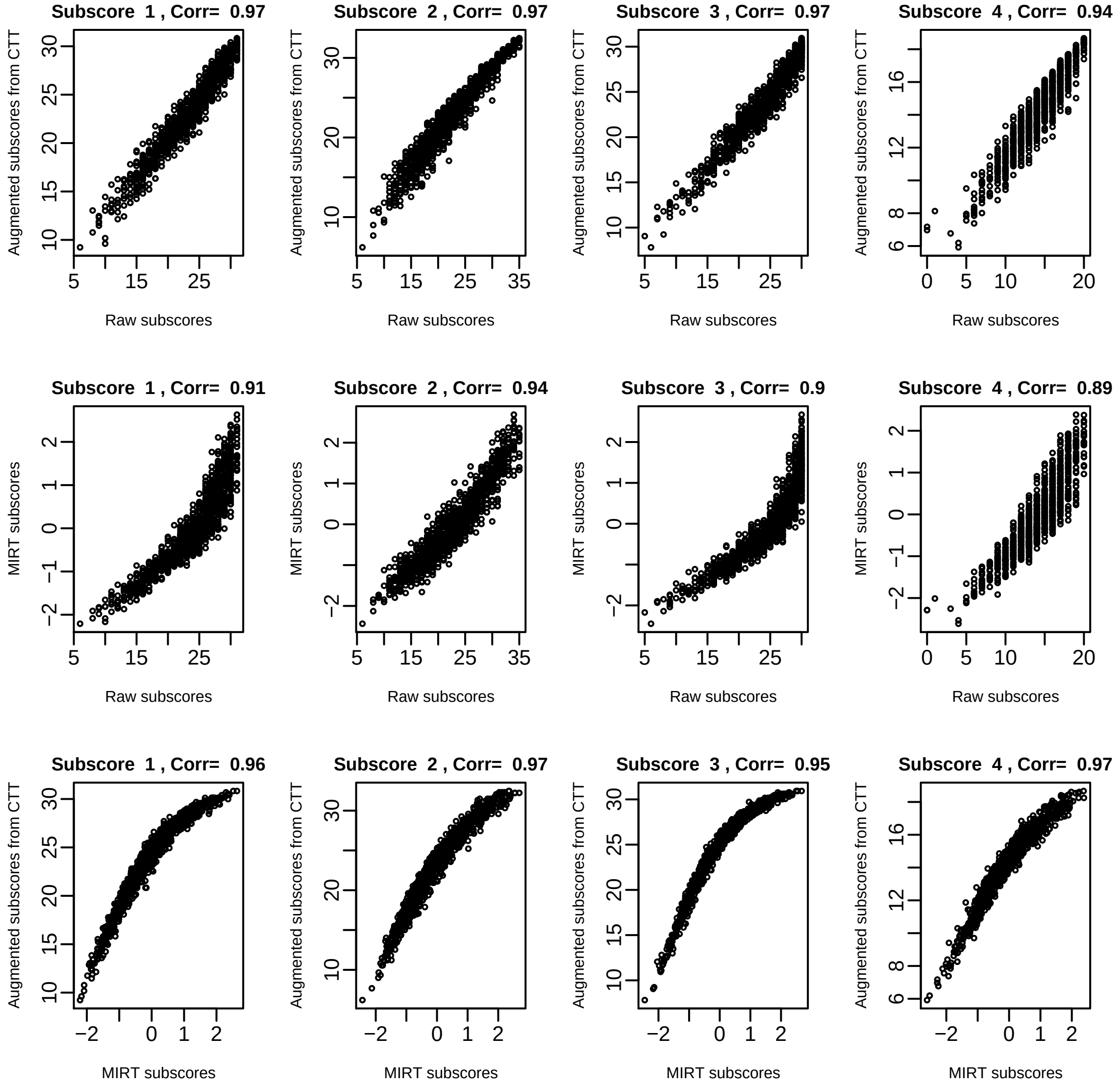

***Figure* 1 Plots of the raw subscores, augmented subscores, and MIRT subscores versus each other for 1,000 examinees for Test C. The correlation coefficients between the variables plotted are also shown.**

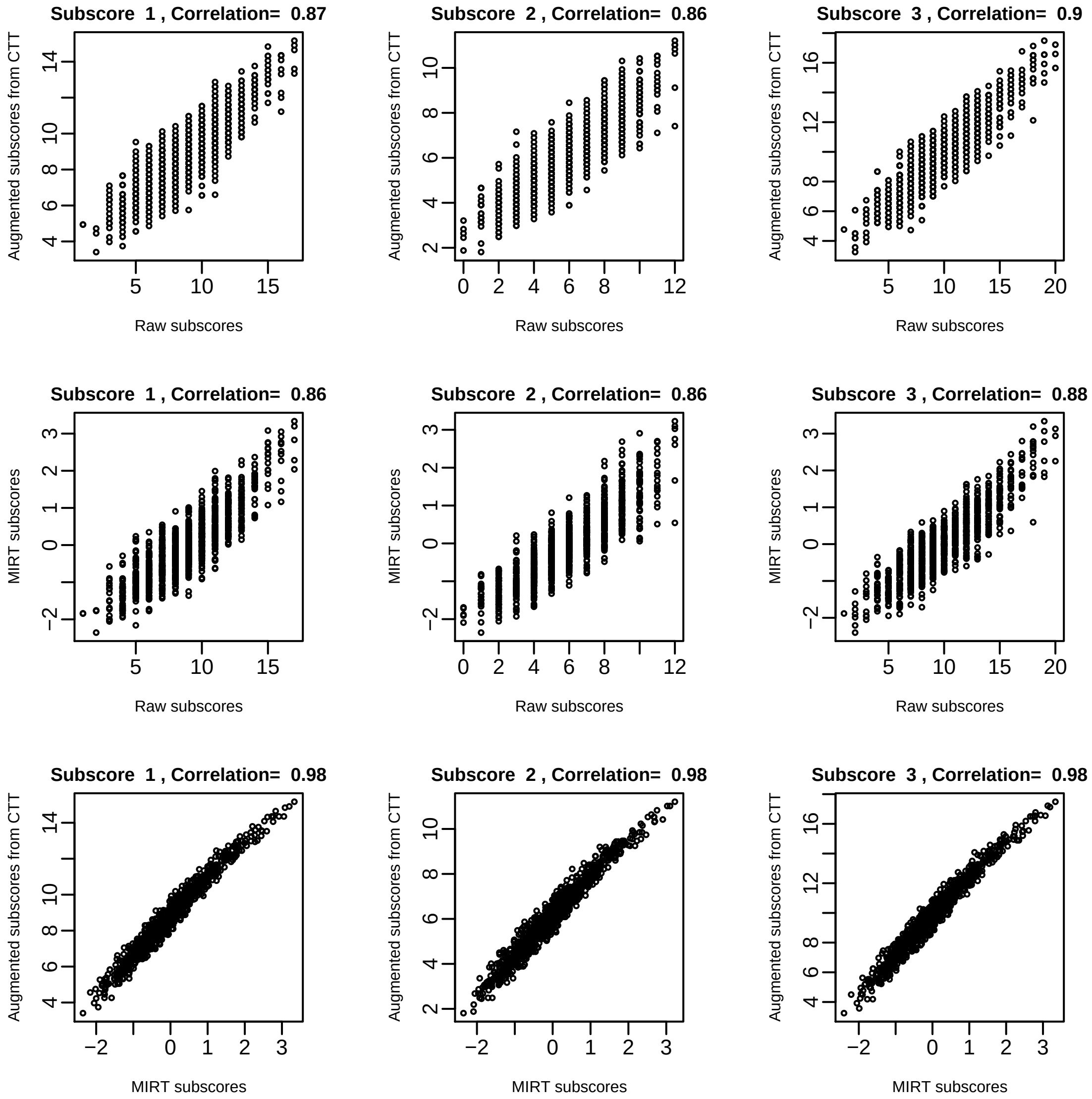

Figure 2 Plots of the raw subscores, augmented subscores, and MIRT subscores versus each other for 1,000 examinees for Test D. The correlation coefficients between the variables plotted are also shown.

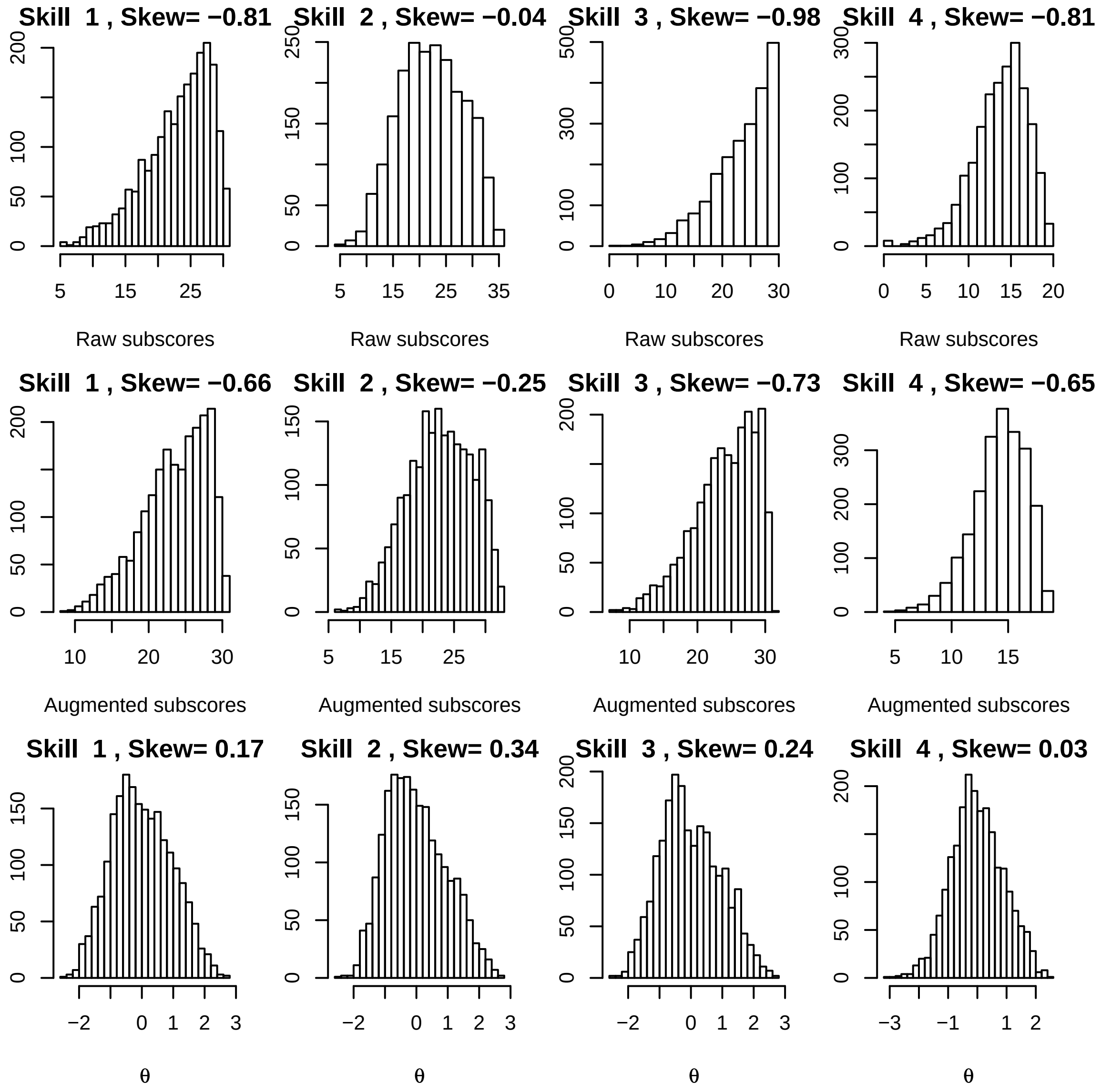

***Figure* 3 Histograms of the distributions of the raw subscores, augmented subscores, and MIRT subscores for Test C. The skewness of the distributions are also shown.**

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295-311.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). *A comparison of subscale score augmentation methods using empirical data.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Fransisco, CA.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187.

Haberman, S. J. (1974). *The analysis of frequency data.* Chicago: University of Chicago Press.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, *18*, 193–211.

Haberman, S. J. (2007). *The information a test provides on an ability parameter* (ETS Research Rep. No. RR-07-18). Princeton, NJ: ETS.

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

Haberman, S. J., Sinharay, S., & Puhan, G. (2006). *Subscores for institutions* (ETS Research Rep. No. RR-06-13). Princeton, NJ: ETS.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (ETS Research Rep. No. RR-08-45). Princeton, NJ: ETS.

Haladyna, S. J., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the health professions*, *24*(7), 349–368.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods* (ETS Research Rep. No. RR-08-54). Princeton, NJ: ETS.

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

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28.

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

Yao, L. H., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83–105.

Yen, W. M. (1987). *A Bayesian/IRT measure of objective performance.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec.

Notes

[1] The augmented subscores also have negative skewness; the MIRT subscores have slightly positive skewness.