PSYCHOMETRIKA—VOL. 75, NO. 2, 209–227
JUNE 2010
DOI: 10.1007/S11336-010-9158-4

REPORTING OF SUBSCORES USING MULTIDIMENSIONAL ITEM RESPONSE THEORY

SHELBY J. HABERMAN AND SANDIP SINHARAY

ETS

Recently, there has been increasing interest in reporting subscores. This paper examines reporting of subscores using multidimensional item response theory (MIRT) models (e.g., Reckase in Appl. Psychol. Meas. 21:25–36, 1997; C.R. Rao and S. Sinharay (Eds), Handbook of Statistics, vol. 26, pp. 607–642, North-Holland, Amsterdam, 2007; Beguin & Glas in Psychometrika, 66:471–488, 2001). A MIRT model is fitted using a stabilized Newton–Raphson algorithm (Haberman in The Analysis of Frequency Data, University of Chicago Press, Chicago, 1974; Sociol. Methodol. 18:193–211, 1988) with adaptive Gauss–Hermite quadrature (Haberman, von Davier, & Lee in ETS Research Rep. No. RR-08-45, ETS, Princeton, 2008). A new statistical approach is proposed to assess when subscores using the MIRT model have any added value over (i) the total score or (ii) subscores based on classical test theory (Haberman in J. Educ. Behav. Stat. 33:204–229, 2008; Haberman, Sinharay, & Puhan in Br. J. Math. Stat. Psychol. 62:79–95, 2008). 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.

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 and Kramer, 2004).

Multidimensional item response theory (MIRT) models (e.g., Reckase, 1997, 2007; Beguin and Glas, 2001) can be employed to report subscores. For instance, de la Torre and Patz (2005) applied a 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, Vevea, Camacho, Reeve, Rosa, & Nelson, 2001). Yao and Boughton (2007) also examined subscore reporting based on a MIRT model and the Markov-chain Monte Carlo (MCMC) algorithm. However, the current approaches of reporting subscores using MIRT models are somewhat problematic in terms of practical application to testing programs with limited time for analysis. For example, 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. Existing software packages that fit MIRT models (mostly using least squares estimation or maximum likelihood estimation) have been used for subscore reporting. See, for example, Ackerman and Shu (2009). However, according to our knowledge, such applications have not considered more than two dimensions. Researchers have also compared different approaches, including the

Note: Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

Requests for reprints should be sent to Sandip Sinharay, ETS, Princeton, NJ, USA. E-mail: ssinharay@ets.org

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), augmentation of raw scores (Wainer et al., 2001), 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 (i) the total score and (ii) subscores based on classical test theory. This work extends to MIRT models the research of Haberman (2008) and Haberman, Sinharay, and Puhan (2008), who suggested methods based on classical test theory (CTT) to examine whether subscores provide any added value over total scores.

Section 1 provides a brief overview of the CTT-based methods of Haberman (2008) and Haberman et al. (2008). 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 Based on Classical Test Theory

This section describes the approach of Haberman (2008) and Haberman et al. (2008) to determine whether and how to report subscores based on CTT. 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 elements 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 $r \geq 1$ skills numbered from 1 to r. Each item j, $1 \leq j \leq q$, intends to measure a single skill (in other words, the data has *simple structure*).[1] It is assumed that each skill corresponds to more than one item.

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

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

and raw subscores

$$S_{ik} = \sum_{j \in J(k)} X_{ij}, \quad 1 \leq k \leq r,$$

where $J(k)$ is the set of items that measure skill k. If $J(k)$ has $q(k)$ members, then S_{ik} has range from 0 to $q(k)$. The true score corresponding to S_i is the true total raw score T_i (which is the

[1] Note that this is not an overly restrictive assumption; in a recent survey of operational tests that report or intend to report subscores, Sinharay (2010) found that each item measures only one skill for more than 20 tests.

average of a large number of observed scores earned by the ith examinee in repeated testings on the same test or on parallel forms of the test), 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 $E([T_{ik} - U_{ikc}]^2)$.

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 $E([T_{ik} - U_{ika}]^2)$. The linear combination U_{ikc} is a special case of U_{ika}. We will often refer to the procedure by which U_{ikc} is obtained as the Haberman augmentation. However, U_{ikc} typically provides results that are very similar to those of U_{ika} and is simpler to compute than U_{ika}. The mean squared errors of U_{ika} and U_{ikc} are almost always equal up to two decimal places (see, e.g., an extensive survey of operational data sets in Sinharay, 2010) and the correlation between U_{ika} and U_{ikc} is almost always larger than 0.99. Hence, we do not provide any results for U_{ika} in this paper.

To compare the possible subscores, this paper employs a measure referred to as proportional reduction in mean squared error (PRMSE). Let τ^2_{k0} be the variance of the true raw subscore T_{ik}, so that τ^2_{k0} is the minimum of $E([T_{ik} - a]^2)$, where a is a constant (the minimum is achieved for $a = E(T_{ik})$). Then τ^2_{ks}, τ^2_{kx}, and τ^2_{kc} cannot exceed τ^2_{k0}. The proportional reductions of mean squared error for the subscores under study are

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

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

and

$$\text{PRMSE}_{kc} = 1 - \tau^2_{kc}/\tau^2_{k0}.$$

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 $\text{PRMSE}_{ks} \leq \text{PRMSE}_{kc}$, and $\text{PRMSE}_{kx} \leq \text{PRMSE}_{kc}$.

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

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 reliability coefficients. For large samples, the decrease in PRMSE due to estimation is negligible.

2. Methods Based on Multivariate Item Response Theory

2.1. *The 2PL MIRT Model*

This paper employs the two-parameter logistic (2PL) MIRT model (Haberman et al., 2008; Reckase, 1997, 2007). The model assumes that an r-dimensional random ability vector $\boldsymbol{\theta}_i$ with elements 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 $\boldsymbol{\theta}_i$. This paper deals with dichotomous items only so that the response X_{ij} can take values of 1 (for a correct answer to an item) and 0 (incorrect answer) only.

As in customary presentations of the compensatory MIRT model (Reckase, 2007), let d_j and $\mathbf{a}_j$ be the real item intercept and r-dimensional item-discrimination vector, respectively, of item j, $1 \le j \le q$. The kth element of $\mathbf{a}_j$, $1 \le k \le r$, denoted as a_{jk}, corresponds to the discrimination of item j with respect to skill k. Given that an examinee has ability vector $\boldsymbol{\theta}$,

$$P_j(1|\boldsymbol{\theta}) = \frac{\exp(\mathbf{a}'_j\boldsymbol{\theta} + d_j)}{1 + \exp(\mathbf{a}'_j\boldsymbol{\theta} + d_j)} \tag{1}$$

is the conditional probability that $X_{ij} = 1$ and

$$P_j(0|\boldsymbol{\theta}) = 1 - P(1|\boldsymbol{\theta}) = \frac{1}{1 + \exp(\mathbf{a}'_j\boldsymbol{\theta} + d_j)} \tag{2}$$

is the conditional probability that $X_{ij} = 0$. In (1) and (2), the vector product

$$\mathbf{a}'_j\boldsymbol{\theta} = \sum_{k=1}^{r} a_{jk}\theta_k.$$

The between-item model (Adams, Wilson, & Wang, 1997) is considered here in which $a_{jk} = 0$ unless j is in $J(k)$. Given examinee ability vector $\boldsymbol{\theta}$,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{q} P_j(x_j|\boldsymbol{\theta}) \tag{3}$$

is the conditional probability that the response vector $\mathbf{X}_i$ is equal to a vector $\mathbf{x}$ with elements x_j, $1 \le j \le q$, each of which is equal to either 0 or 1. The unconditional probability that $\mathbf{X}_i = \mathbf{x}$ is then

$$p(\mathbf{x}) = E\big(p(\mathbf{x}|\boldsymbol{\theta}_i)\big). \tag{4}$$

In (4), $p(\mathbf{x}|\boldsymbol{\theta}_i)$ is the random variable defined so that $p(\mathbf{x}|\boldsymbol{\theta}_i)$ is equal to $p(\mathbf{x}|\boldsymbol{\theta})$ when the random ability vector $\boldsymbol{\theta}_i$ of examinee i is equal to $\boldsymbol{\theta}$. If $\boldsymbol{\theta}_i$ is a continuous random vector with density f, then

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{5}$$

2.2. *Estimation for the 2PL MIRT Model*

In this paper, maximum marginal likelihood is applied under the model that $\boldsymbol{\theta}_i$ has a multivariate normal distribution $N(\mathbf{0}, \mathbf{D})$. Here, $\mathbf{0}$ is the r-dimensional vector with all elements 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 under the between-item model of the item parameters a_{jk} and d_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). It appears that the practical effect of the multivariate normality assumption is minor. In addition, the analysis does not assume that the model used is actually correct.

Computation of maximum marginal likelihood estimates of the model parameters $\mathbf{a}_j$, $1 \le j \le q$, d_j, $1 \le j \le q$, and $D_{kk'}$, $1 \le k < k' \le r$, may be performed by a version of the stabilized Newton–Raphson algorithm (Haberman, 1988) described in Haberman et al. (2008). Because calculations employ adaptive multivariate Gauss–Hermite integration (Schilling & Bock, 2005), computational time is not excessive. The software programs used in this paper are available on request from the authors.

If the MIRT model holds, then the model-based probability $p(\mathbf{x})$ given by (5) matches the actual (and unknown) probability distribution of each observed vector $\mathbf{X}_i$ of responses. If the model does not hold, then the maximum likelihood method estimates parameters that provide the best correspondence between $p(\mathbf{x})$ given by (5) and the actual probability distribution of each vector $\mathbf{X}_i$. In addition, even if the model does not hold, the procedure of Haberman (2007) can be used for each examinee i to construct random vectors $\boldsymbol{\theta}_i^*$ that approximate the ability vector $\boldsymbol{\theta}_i$; in this procedure, information theory is used to construct $\boldsymbol{\theta}_i^*$ with conditional distribution given $\mathbf{X}_i$ that is of the same form as the conditional distribution of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i$ if the model holds. For further details, see Haberman and Sinharay (2010).

2.3. *Subscores Based on MIRT (The Expected a Posteriori Means)*

Given that $\mathbf{X}_i = \mathbf{x}$, the conditional density of $\boldsymbol{\theta}_i$ has value

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

at the r-dimensional vector $\boldsymbol{\theta}$. The conditional expected value of $\boldsymbol{\theta}_i$ is

$$E(\boldsymbol{\theta}_i|\mathbf{x}) = \int \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

Consider the random variable $f(\boldsymbol{\theta}|\mathbf{X}_i)$ with value $f(\boldsymbol{\theta}|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$. The unconditional density of $\boldsymbol{\theta}_i$ at $\boldsymbol{\theta}$ is then the expected value

$$g(\boldsymbol{\theta}) = E\big(f(\boldsymbol{\theta}|\mathbf{X}_i)\big) = \sum_{\mathbf{x} \in \Gamma} f(\boldsymbol{\theta}|\mathbf{x}) P(\mathbf{X}_i = \mathbf{x}) \tag{6}$$

of $f(\boldsymbol{\theta}|\mathbf{X}_i)$.

Let $\tilde{\boldsymbol{\theta}}_i$ be the random vector $E(\boldsymbol{\theta}_i|\mathbf{X}_i)$ with value $E(\boldsymbol{\theta}_i|\mathbf{x})$ if $\mathbf{X}_i=\mathbf{x}$. Then $\tilde{\boldsymbol{\theta}}_i$, 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 MIRT models. Exchange of integration and summation and application of (6) shows that $\tilde{\boldsymbol{\theta}}_i$ has expectation

$$E(\tilde{\boldsymbol{\theta}}_i)=\sum_{\mathbf{x}\in\Gamma}P(\mathbf{X}_i=\mathbf{x})\int\boldsymbol{\theta}f(\boldsymbol{\theta}|\mathbf{x})\,d\boldsymbol{\theta}=\int\boldsymbol{\theta}g(\boldsymbol{\theta})\,d\boldsymbol{\theta}=E(\boldsymbol{\theta}_i).$$

The covariance matrix of $\boldsymbol{\theta}_i$ is

$$\operatorname{Cov}(\boldsymbol{\theta}_i)=\int\big[\boldsymbol{\theta}-E(\boldsymbol{\theta}_i)\big]\big[\boldsymbol{\theta}-E(\boldsymbol{\theta}_i)\big]'g(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

Here, the prime indicates a transpose, and it should be noted that $E(\boldsymbol{\theta}_i)$ is a constant vector. Given that $\mathbf{X}_i=\mathbf{x}$, the approximation error $\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i$ has conditional mean $\mathbf{0}$ and conditional covariance matrix

$$\operatorname{Cov}(\boldsymbol{\theta}_i|\mathbf{x})=\int\big[\boldsymbol{\theta}-E(\boldsymbol{\theta}_i|\mathbf{x})\big]\big[\boldsymbol{\theta}-E(\boldsymbol{\theta}_i|\mathbf{x})\big]'f(\boldsymbol{\theta}|\mathbf{x})\,d\boldsymbol{\theta}.$$

If $\operatorname{Cov}(\boldsymbol{\theta}_i|\mathbf{X}_i)$ is the random matrix with value $\operatorname{Cov}(\boldsymbol{\theta}_i|\mathbf{x})$ when $\mathbf{X}_i=\mathbf{x}$, then $\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i$ has unconditional mean $\mathbf{0}$ and unconditional covariance matrix,

$$\operatorname{Cov}(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)=E\big(\operatorname{Cov}(\boldsymbol{\theta}_i|\mathbf{X}_i)\big)=\sum_{\mathbf{x}\in\Gamma}P(\mathbf{X}_i=\mathbf{x})\operatorname{Cov}(\boldsymbol{\theta}_i|\mathbf{x}).\tag{7}$$

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

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

for $1\le k'\le r$. The kth element θ_{ik} of $\boldsymbol{\theta}_i$ has variance $\tau^2_{k0\theta}=\boldsymbol{\delta}'_k\operatorname{Cov}(\boldsymbol{\theta}_i)\boldsymbol{\delta}_k$. Equation 7 implies that the mean squared error $\tau^2_{k\theta}$ for the kth element $\tilde{\theta}_{ik}$ of $\tilde{\boldsymbol{\theta}}_i$ is

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

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

2.4. *PRMSEs of the Subscores Based on MIRT*

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

$$\rho^2(\mathbf{c})=1-\frac{\mathbf{c}'\operatorname{Cov}(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\mathbf{c}}{\mathbf{c}'\operatorname{Cov}(\boldsymbol{\theta}_i)\mathbf{c}}.\tag{8}$$

The quantity $\mathbf{c}'\operatorname{Cov}(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\mathbf{c}$ in (8) 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, $\mathbf{c}'\operatorname{Cov}(\boldsymbol{\theta}_i)\mathbf{c}$ in (8) 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., 2008).

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

$$\text{PRMSE}_{k\theta M} = \rho^2(\boldsymbol{\delta}_k) = 1 - \tau^2_{k\theta}/\tau^2_{k0\theta}.$$

2.5. *Estimation of EAP Means and PRMSEs*

The EAP mean $\tilde{\boldsymbol{\theta}}_i$ depends on unknown parameters; however, parameter estimates are available. Let $\hat{f}$ be the density of a random variable with a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\hat{\mathbf{D}}$. Let

$$\hat{P}_j(1|\boldsymbol{\theta}) = \frac{\exp(\hat{\mathbf{a}}'_j\boldsymbol{\theta} + \hat{d}_j)}{1 + \exp(\hat{\mathbf{a}}'_j\boldsymbol{\theta} + \hat{d}_j)},$$

$$\hat{P}_j(0|\boldsymbol{\theta}) = 1 - \hat{P}(1|\boldsymbol{\theta}),$$

$$\hat{p}(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{q} \hat{P}_j(x_j|\boldsymbol{\theta}),$$

$$\hat{p}(\mathbf{x}) = \int \hat{p}(\mathbf{x}|\boldsymbol{\theta})\hat{f}(\boldsymbol{\theta})\,d\boldsymbol{\theta},$$

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

and

$$\hat{E}(\boldsymbol{\theta}|\mathbf{x}) = \int \boldsymbol{\theta}\hat{f}(\boldsymbol{\theta}|\mathbf{x})\,d\boldsymbol{\theta}.$$

Let $\hat{\boldsymbol{\theta}}_i$ be the random vector with value $\hat{E}(\boldsymbol{\theta}|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$. Then $\hat{\boldsymbol{\theta}}_i$ may be employed in place of $\tilde{\boldsymbol{\theta}}_i$.

To study reliability, further estimates are required. Let

$$\widehat{\text{Cov}}(\boldsymbol{\theta}|\mathbf{x}) = \int \big[\boldsymbol{\theta} - \hat{E}(\boldsymbol{\theta}|\mathbf{x})\big]\big[\boldsymbol{\theta} - \hat{E}(\boldsymbol{\theta}|\mathbf{x})\big]'\hat{f}(\boldsymbol{\theta}|\mathbf{x})\,d\boldsymbol{\theta},$$

let $\widehat{\text{Cov}}(\boldsymbol{\theta}|\mathbf{X}_i)$ be the random matrix with value $\widehat{\text{Cov}}(\boldsymbol{\theta}|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$, and let

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

The quantity on the right-hand side of the above equation is the average over the sample of the estimated posterior variance matrix of the examinees. Let $\hat{f}(\boldsymbol{\theta}|\mathbf{X}_i)$ be the random variable with value $\hat{f}(\boldsymbol{\theta}|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$, and let

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

Let

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

and let

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

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

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

The estimated variance of $\hat{\theta}_{ik}$, element k of $\hat{\boldsymbol{\theta}}_{ik}$, is given by $\hat{\tau}^2_{k0\theta} = \boldsymbol{\delta}'_k\widehat{\mathrm{Cov}}(\boldsymbol{\theta})\boldsymbol{\delta}_k$, and for the kth element $\tilde{\theta}_{ik}$ of $\tilde{\boldsymbol{\theta}}_i$, the estimated mean squared error $\hat{\tau}^2_{k\theta}$ for approximation of θ_{ik} by $\hat{\theta}_{ik}$ is $\boldsymbol{\delta}'_k\widehat{\mathrm{Cov}}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\boldsymbol{\delta}_k$. The estimated PRMSE is $\widehat{\mathrm{PRMSE}}_{k\theta M} = 1 - \hat{\tau}^2_{k\theta}/\hat{\tau}^2_{k0\theta}$. The reliability index for each of the dimensions described in Adams et al. (1997) is similar in spirit to this estimated PRMSE.

For each skill k, we also fitted a unidimensional IRT (UIRT) model for items i in $J(k)$. For these items, the MIRT model applies with $r = 1$. Equivalently, the r unidimensional IRT models yield the same result as the r-dimensional MIRT model in which the added restriction is imposed that $D_{kk'} = 0$ for $k \neq k'$. The estimated PRMSE obtained from the univariate computations is denoted by $\widehat{\mathrm{PRMSE}}_{k\theta U}$. The estimated marginal reliability described in, for example, Thissen, Nelson, and Swygert (2001), is similar in spirit to this estimated PRMSE.

2.6. *An Alternative Approach for Reporting MIRT-based Subscores: The Use of the Test Characteristic Curves*

The reader may be concerned about comparison of PRMSE for approximation of an element of $\boldsymbol{\theta}_i$ (or $\mathrm{PRMSE}_{k\theta M}$) to PRMSE for approximation of a true subscore T_{ik} (PRMSE_{kc}) because the θ_{ik} do not appear to be on the same scale as the true scores T_{ik}. Several approaches to this issue can be considered. One approach just notes that PRMSE measures are dimensionless. The interest should be in metrics for measurement of performance in which PRMSE is largest without regard to scale.

An alternative approach considers transformation of θ_{ik} into the same scale as the raw subscore. For this purpose, the test characteristic curves for the S_{ik} (Hambleton, Swaminathan, & Rogers, 1991, p. 85) can be employed. Let the test characteristic curve for skill k be

$$V_k(\boldsymbol{\theta}) = \sum_{j\in J(k)} P_j(1|\boldsymbol{\theta}).$$

Under the model, the true score T_{ik} is $V_k(\boldsymbol{\theta}_i)$, and $V_k(\boldsymbol{\theta}_i)$ is a function of θ_{ik}. In general, $V_k(\boldsymbol{\theta}_i)$ has possible values from 0 to $q(k)$, the number of elements of $J(k)$. Note that the raw subscore S_{ik} also satisfies $0 \le S_{ik} \le q(k)$. Consider approximation of T_{ik} by

$$U_{ik\theta} = E\big(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i\big).$$

Here, $E(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i)$ is $E(V_k(\boldsymbol{\theta}_i)|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$ and

$$E\big(V_k(\boldsymbol{\theta}_i)|\mathbf{x}\big) = \int V_k(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

In addition, $E(V_k(\boldsymbol{\theta}_i)|\mathbf{x})$ is also the conditional expected value of $V_k(\boldsymbol{\theta}_i^*)$ given that $\mathbf{X}_i^* = \mathbf{x}$. Then $U_{ik\theta}$ is the expected *a posteriori* mean of $V_k(\boldsymbol{\theta}_i)$ given $\mathbf{X}_i$. The expectation of $U_{ik\theta}$ is

$$E(U_{ik\theta}) = \sum_{\mathbf{x}\in\Gamma} P(\mathbf{X}_i = \mathbf{x}) \int V_k(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = \int V_k(\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E\big(V_k(\boldsymbol{\theta}_i)\big).$$

The variance of $V_k(\boldsymbol{\theta}_i)$ is

$$\sigma^2\big(V_k(\boldsymbol{\theta}_i)\big) = \int \big[V_k(\boldsymbol{\theta}) - E\big(V_k(\boldsymbol{\theta}_i)\big)\big]^2 g(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Given that $\mathbf{X}_i = \mathbf{x}$, the approximation error $U_{ik\theta} - V_k(\boldsymbol{\theta}_i)$ has conditional mean 0 and conditional variance

$$\sigma^2\big(V_k(\boldsymbol{\theta}_i)|\mathbf{x}\big) = \int \big[V_k(\boldsymbol{\theta}) - E\big(V_k(\boldsymbol{\theta}_i)|\mathbf{x}\big)\big]^2 f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

If $\sigma^2(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i)$ is the random variable with value $\sigma^2(V_k(\boldsymbol{\theta}_i)|\mathbf{x})$ when $\mathbf{X}_i = \mathbf{x}$, then $U_{ik\theta} - V_k(\boldsymbol{\theta}_i)$ has unconditional mean 0 and unconditional variance

$$\sigma^2\big(U_{ik\theta} - V_k(\boldsymbol{\theta}_i)\big) = E\big(\sigma^2\big(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i\big)\big) = \sum_{\mathbf{x}\in\Gamma} P(\mathbf{X}_i = \mathbf{x})\sigma^2\big(V_k(\boldsymbol{\theta}_i)|\mathbf{x}\big).$$

The reliability of $U_{ik\theta}$ is then

$$1 - \frac{\text{Error variance}}{\text{Total variance}} = 1 - \frac{E(\sigma^2(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i)}{\sigma^2(V_k(\boldsymbol{\theta}_i))}.$$

This reliability may be denoted by $\text{PRMSE}_{Vk\theta M}$.

As in the case of $\tilde{\boldsymbol{\theta}}_i$, parameter estimates must be employed in practice. One may let

$$\hat{E}\big(V_k(\boldsymbol{\theta})|\mathbf{x}\big) = \int V_k(\boldsymbol{\theta}) \hat{f}(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta},$$

and let $\hat{U}_{ik\theta}$ be the random variable with value $\hat{E}(V_k(\boldsymbol{\theta})|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$. To estimate reliability, let

$$\hat{\sigma}^2\big(V_k(\boldsymbol{\theta})|\mathbf{x}\big) = \int \big[V_k(\boldsymbol{\theta}) - \hat{E}\big(V_k(\boldsymbol{\theta})|\mathbf{x}\big)\big]^2 \hat{f}(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta},$$

let $\hat{\sigma}^2(V_k(\boldsymbol{\theta})|\mathbf{X}_i)$ be the random variable with value $\hat{\sigma}^2(V_k(\boldsymbol{\theta})|\mathbf{x})$ if $\mathbf{X}_i = \mathbf{x}$, and let

$$\hat{\sigma}^2\big(U_{ik\theta} - V_k(\boldsymbol{\theta}_i)\big) = n^{-1}\sum_{i=1}^{n} \hat{\sigma}^2\big(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i\big).$$

Let

$$\bar{U}_{k\theta} = \int V_k(\boldsymbol{\theta}) \hat{g}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = n^{-1}\sum_{i=1}^{n} \hat{U}_{ik\theta},$$

and let

$$\hat{\sigma}^2\big(V_k(\boldsymbol{\theta})\big) = \int \big[V_k(\boldsymbol{\theta}) - \bar{U}_{k\theta}\big]^2 \hat{g}(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where adaptive quadrature can be used to compute the integral. For large samples, the reliability for $\hat{U}_{ik}$ is approximated by

$$\widehat{\text{PRMSE}}_{Vk\theta M} = 1 - \hat{\sigma}^2\big(U_{ik\theta} - V_k(\boldsymbol{\theta}_i)\big) / \hat{\sigma}^2\big(V_k(\boldsymbol{\theta})\big).$$

Similar calculations are possible when a unidimensional IRT model is fitted for each skill. The estimated reliability in this case for skill k for the EAP estimate of the test characteristic curve $V_k(\boldsymbol{\theta}_i)$ will be denoted by $\widehat{\text{PRMSE}}_{Vk\theta U}$.

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 find out the best possible way to report subscores for these tests. To fit the MIRT model given by (3) and compute the PRMSEs, 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, 2, 3, 4, 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 (2010). The reliability of the different scores and subscores were estimated using Cronbach's α.

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 is composed of 116 MC questions divided 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.

TABLE 1.
Results for Test A.

	Subscores			
	1	2	3	4
Length	30	29	30	30
Mean subscore	22.9	18.9	17.6	19.1
Standard deviation of subscore	3.8	5.4	4.5	4.3
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
Estimated 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$_{k\theta U}$	0.71	0.83	0.77	0.75
PRMSE$_{Vk\theta U}$	0.75	0.85	0.77	0.76
PRMSE$_{k\theta M}$	0.84	0.87	0.85	0.87
PRMSE$_{Vk\theta M}$	0.86	0.88	0.85	0.88

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. Copyright by Educational Testing Service, All rights reserved.

TABLE 2.
Results for Test B.

	Subscores		
	1	2	3
Length	13	10	33
Mean subscore	8.1	7.1	23.3
Standard deviation of subscore	2.1	1.5	3.7
Correlation between the raw subscores	1.00	0.34	0.51
	0.96	1.00	0.41
	0.95	**0.99**	1.00
Estimated 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$_{k\theta U}$	0.49	0.32	0.65
PRMSE$_{Vk\theta U}$	0.51	0.36	0.68
PRMSE$_{k\theta M}$	0.74	0.71	0.75
PRMSE$_{Vk\theta M}$	0.76	0.73	0.77

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. Copyright by Educational Testing Service, All rights reserved.

Table 3.
Results for Test C.

	Subscores			
	1	2	3	4
Length	31	35	30	20
Mean subscore	23.6	22.8	24.1	14.2
Standard deviation of subscore	5.1	6.0	5.1	3.2
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
Estimated 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
PRMSE$_{k\theta U}$	0.82	0.85	0.82	0.68
PRMSE$_{Vk\theta U}$	0.86	0.86	0.88	0.72
PRMSE$_{k\theta M}$	0.90	0.90	0.91	0.79
PRMSE$_{Vk\theta M}$	0.91	0.91	0.93	0.79

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. Copyright by Educational Testing Service, All rights reserved.

Table 4.
Results for Test D.

	Subscores		
	1	2	3
Length	17	12	20
Mean subscore	8.8	6.1	9.8
Standard deviation of subscore	2.8	2.4	3.3
Correlation between the raw subscores	1.00	0.57	0.61
	0.95	1.00	0.58
	0.97	**0.94**	1.00
Estimated 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
PRMSE$_{k\theta U}$	0.65	0.66	0.70
PRMSE$_{Vk\theta U}$	0.65	0.68	0.70
PRMSE$_{k\theta M}$	0.83	0.81	0.84
PRMSE$_{Vk\theta M}$	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. Copyright by Educational Testing Service, All rights reserved.

TABLE 5.
Results for Test E.

	Subscores		
	1	2	3
Length	25	23	25
Mean subscore	18.3	15.0	18.6
Standard deviation of subscore	5.2	4.7	4.9
Correlation between the raw subscores	1.00	0.76	0.79
	0.90	1.00	0.73
	0.91	**0.86**	1.00
Estimated 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
PRMSE$_{k\theta U}$	0.84	0.84	0.82
PRMSE$_{Vk\theta U}$	0.84	0.84	0.82
PRMSE$_{k\theta M}$	0.91	0.89	0.90
PRMSE$_{Vk\theta M}$	0.92	0.90	0.91

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. Copyright by Educational Testing Service, All rights reserved.

Test D is designed to assess the mathematical knowledge and competencies necessary for a beginning teacher of secondary school mathematics. It is composed of 49 MC questions divided 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 (20 items). The sample size was 6,818, and the reliability of the total test score was 0.82.

Test E is used to measure skill necessary for prospective and practicing paraprofessionals. It is composed of 73 MC questions divided 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 to 5 provide results for Tests A to E. Each of these tables shows the following:

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

These tables refer to the subscores as subscores $1, 2, \ldots$ (the descriptions of the subscores are given earlier in the "Data Sets" subsection). Note that a comparison of PRMSE$_{kc}$ with both of PRMSE$_{k\theta M}$ and PRMSE$_{Vk\theta M}$ will reveal if the MIRT approach provides subscores that, relative to their variability, are more accurate than those provided by the CTT approach. A comparison of $\text{PRMSE}_{k\theta M}$ and $\text{PRMSE}_{k\theta U}$ (and also of $\text{PRMSE}_{Vk\theta M}$ and $\text{PRMSE}_{Vk\theta U}$) will reveal how much one gains by employing a MIRT model over a UIRT model. A comparison of PRMSE_{ks} with both of $\text{PRMSE}_{k\theta U}$ and $\text{PRMSE}_{Vk\theta U}$ will reveal how much one gains by using subscores based on a UIRT model rather than using the raw subscores.

The pattern of results is quite consistent. The estimated PRMSE of the MIRT subscores (both $\text{PRMSE}_{Vk\theta M}$ and $\text{PRMSE}_{k\theta M}$) are almost always as high as, or higher than, the estimated PRMSE of the augmented subscores. The differences are often quite small, but they are appreciable in a number of cases. The estimates of $\text{PRMSE}_{Vk\theta M}$ are as high as, or higher than those of $\text{PRMSE}_{k\theta M}$ for all of our data sets.

The estimates of PRMSEs of subscores based on UIRT (both $\text{PRMSE}_{Vk\theta U}$ and $\text{PRMSE}_{k\theta U}$) are mostly slightly higher than the subscore reliability, but are substantially less than the estimated PRMSEs based on MIRT. The estimates of $\text{PRMSE}_{Vk\theta U}$ are as high as, or higher than those of $\text{PRMSE}_{k\theta U}$ for all of our data sets.

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, (i) scatter plots of augmented subscores versus raw subscores (the panels in the top row), (ii) the MIRT subscores versus the raw subscores (the panels in the middle row), and (iii) 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 (denoted as "r") and rank correlations (denoted as "rs") between the variables being plotted. Results were similar for the other tests and are not shown. While the simple correlations between the raw subscores and the augmented/MIRT subscores are between 0.86 and 0.97, the correlation between the MIRT subscores and the augmented subscores are very close to 1. Our finding of the similarity of the MIRT subscores and augmented subscores supports the finding in de la Torre and Patz (2005) of the similarity of MCMC-based MIRT subscores and Wainer-augmented subscores. High correlations do not imply perfectly linear relationships. For example, 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);[2] this is one reason behind the curvilinear relationship in Figure 1. In addition, it should be noted that some nonlinearity is expected because the MIRT scores are not on the same scale as are the original subscores.

Figure 4 shows for each of the 4 subscores of Test C (top row) and for each of the 3 subscores of the Test D (bottom row), scatter plots of raw subscores versus UIRT subscores raw subscores for 1,000 randomly chosen examinees. Each panel also shows the correlation between the variables being plotted. The plots point to the extremely high correlations and rank correlations between the raw subscores and UIRT subscores—that is expected given the closeness of $\text{PRMSE}_{k\theta U}$ and $\text{PRMSE}_{Vk\theta U}$ to PRMSE_{ks} for these tests. Results were similar for the other tests and are not shown here.

The results indicate that Haberman augmentation and the MIRT results strongly dominate the results for estimates 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, the augmented subscores do not lead to any improvement over the total score. The estimated values of $\text{PRMSE}_{k\theta M}$ indicate that the MIRT-based subscores do not lead to any improvement over the total score either. However, the estimated values of $\text{PRMSE}_{Vk\theta M}$ indicate that the MIRT-based subscores lead to some improvement over the total score. This is an example when the differences between $\text{PRMSE}_{k\theta M}$ and $\text{PRMSE}_{Vk\theta M}$ are practically significant.

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

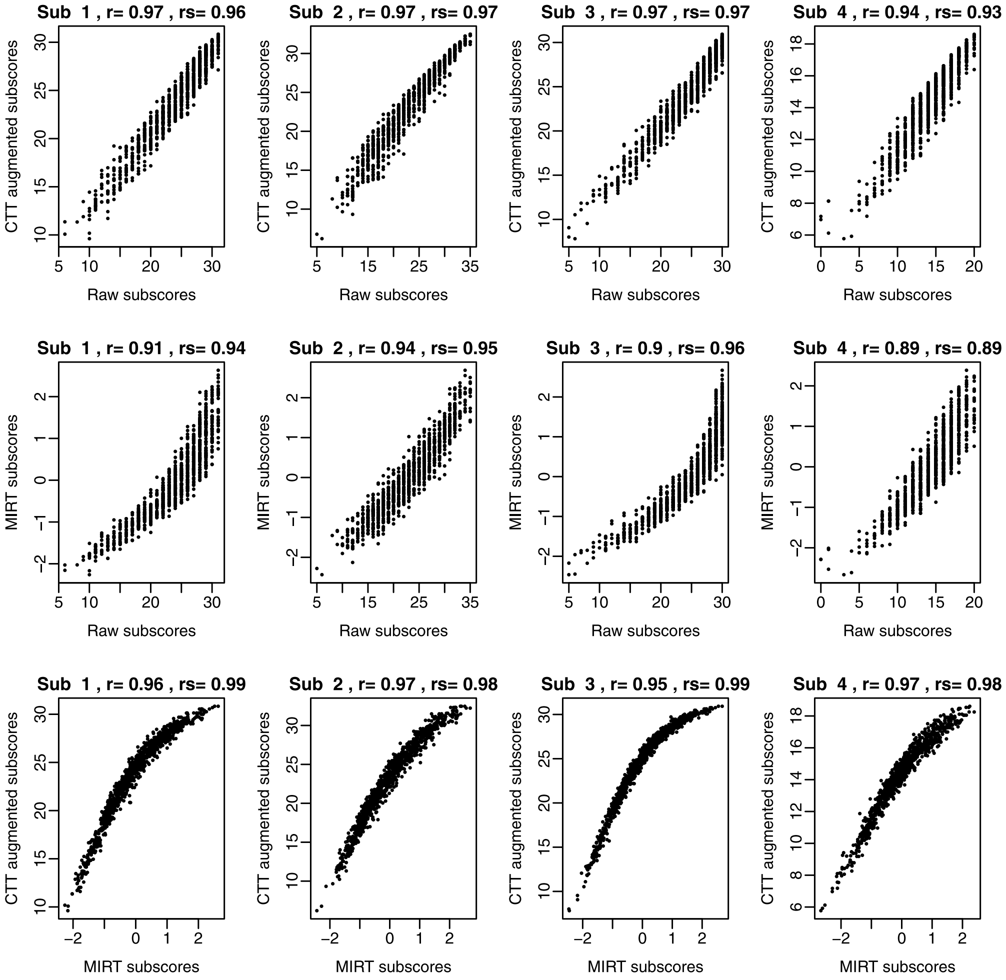

FIGURE 1.
Plots of the raw subscores, augmented subscores, and MIRT subscores versus each other for 1,000 examinees for test C. The correlation coefficients (simple and rank) between the variables plotted are also shown. “Sub” denotes subscore, “r” denotes correlation, and “rs” denotes rank correlation. Copyright by Educational Testing Service, All rights reserved.

4. Conclusions

The use of MIRT models to generate subscores is quite feasible, as evident from the examples. While we have considered a variety of tests and are confident that the types of results observed in this paper will hold in general, it is possible to obtain different results for other samples and/or other tests. Given the similarity of the IRT-based 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 report subscores.

Computational burden for the MIRT analysis appears acceptable—the software program did not take more than a couple of hours for any of the data sets we analyzed here. However, several details of calculation would best be modified for much larger samples. The six quadrature points per dimension was 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

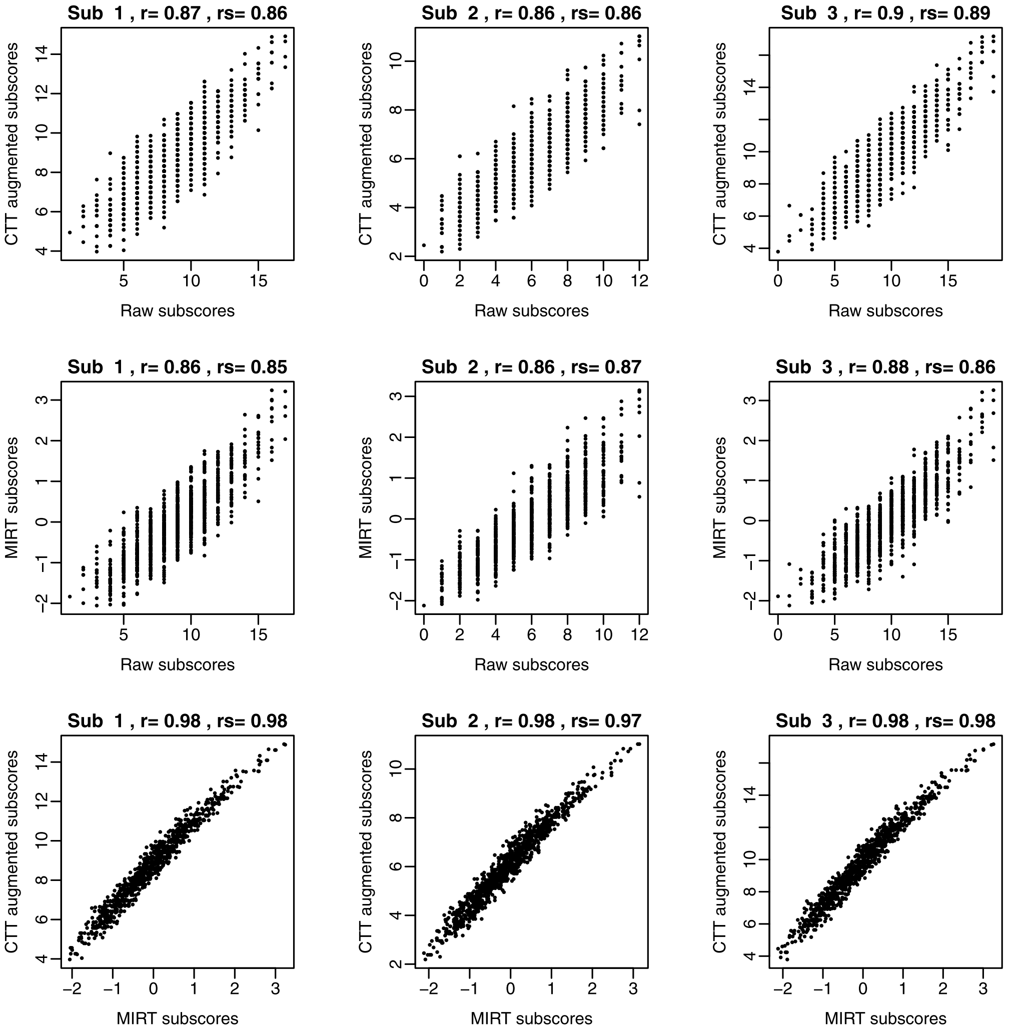

FIGURE 2.
Plots of the raw subscores, augmented subscores, and MIRT subscores versus each other for 1,000 examinees for test D. The correlation coefficients (simple and rank) between the variables plotted are also shown. “Sub” denotes subscore, “r” denotes correlation, and “rs” denotes rank correlation. Copyright by Educational Testing Service, All rights reserved.

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 to 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 compute subscores does involve a much higher computational burden than the CTT-based approach of Haberman (2008) and Haberman et al. (2008).

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 transformations of the $\boldsymbol{\theta}_i$s to the same scale as the raw subscores as discussed in Section 2. For example, instead of reporting the posterior means $E(\boldsymbol{\theta}_i|\mathbf{X}_i)$, one might report $E(V_k(\boldsymbol{\theta}_i)|\mathbf{X}_i)$, which is the posterior mean of the true kth subscore and is on the same scale as the subscore k.

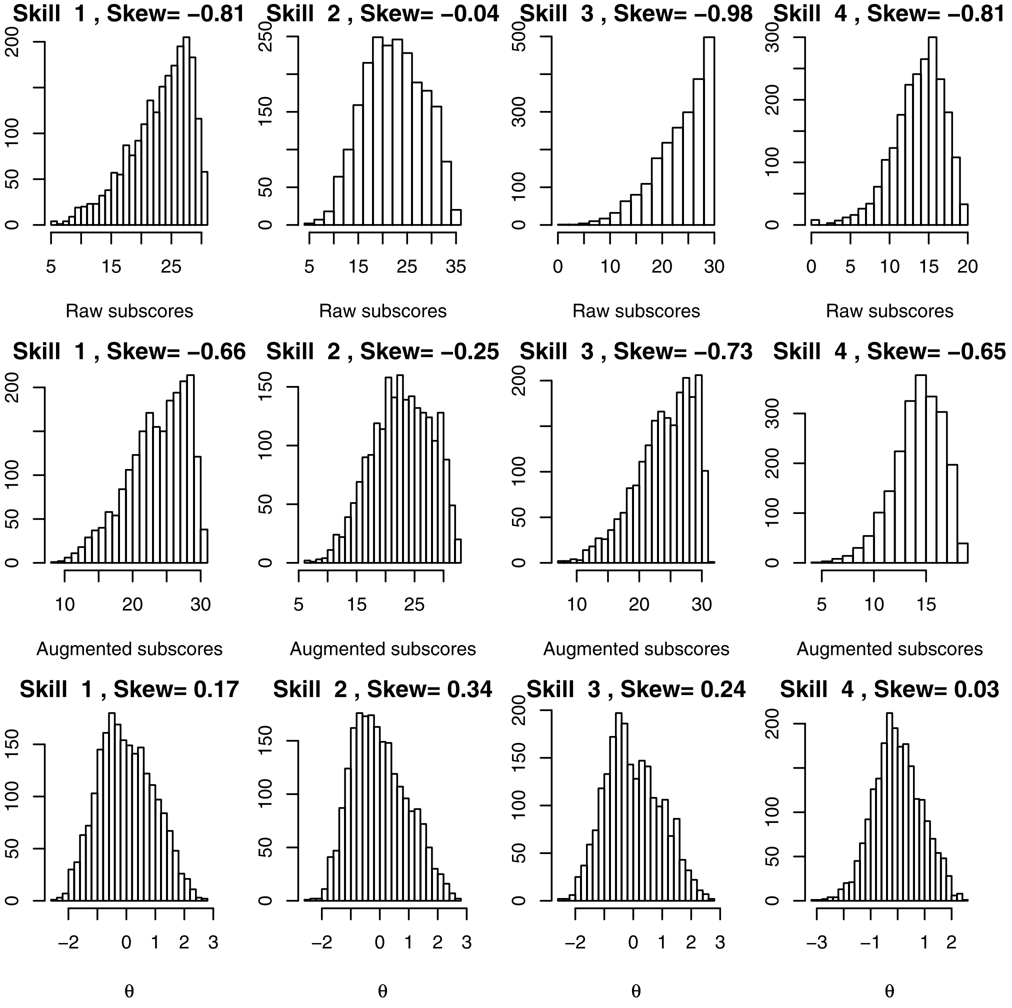

FIGURE 3.
Histograms of the distributions of the raw subscores, augmented subscores, and MIRT subscores for test C. The skewness of the distributions are also shown. Copyright by Educational Testing Service, All rights reserved.

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 in a decision on if 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 that anchor test equating of the subscore is not feasible.

In any work concerning subscores, it is important to emphasize that statistical analysis will not convert educational tests carefully constructed to be unidimensional into tests that yield useful diagnostic scores for many different skills. As Luecht, Gierl, Tan, and Huff (2006) warn us, we should not try to extract something that is not there.

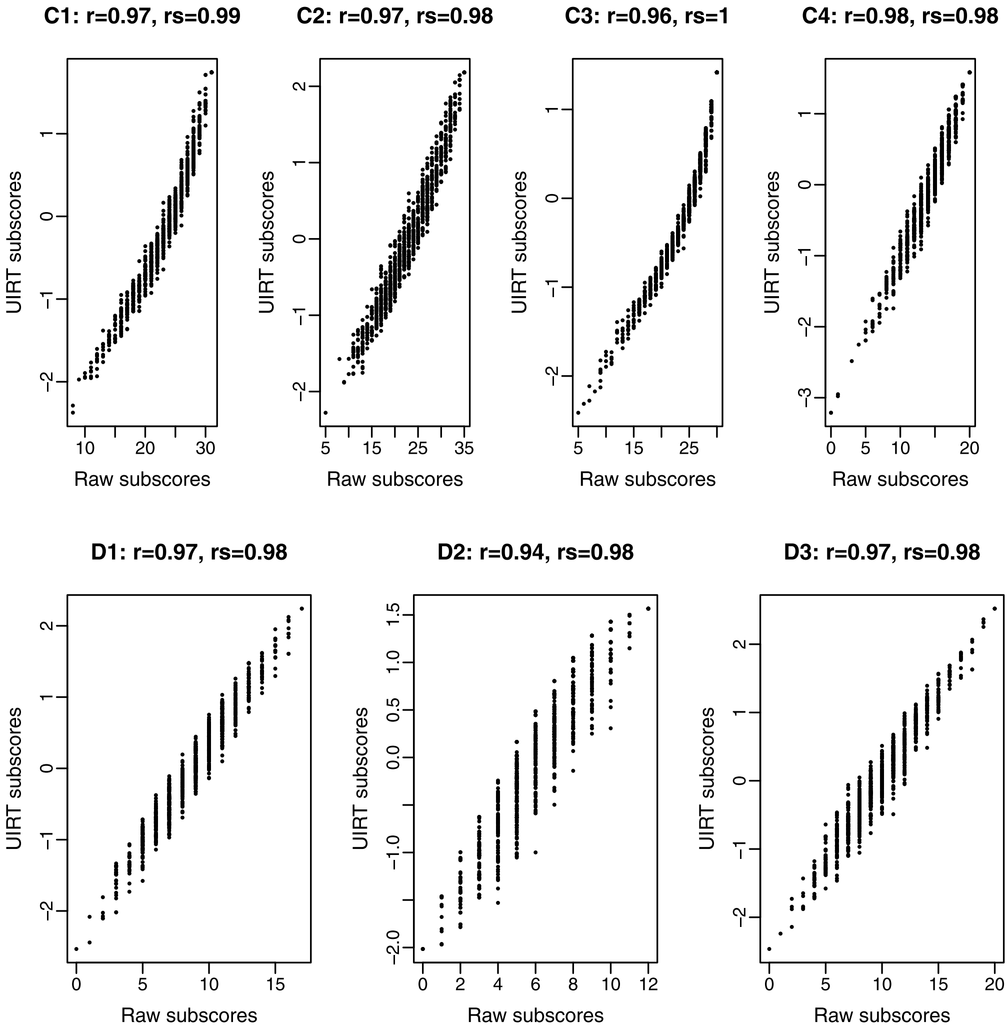

FIGURE 4.
Plots of the raw subscores and UIRT subscores for 1,000 examinees for tests C and D. The correlation coefficients (simple and rank) between the variables plotted are also shown. C1, C2, C3, and C4 denote the four subscores of Test C, D1, D2, D3, and D4 denote the four subscores of Test D, "r" denotes correlation, and "rs" denotes rank correlation. Copyright by Educational Testing Service, All rights reserved.

References

Ackerman, T., & Shu, Z. (2009). *Using confirmatory mirt modeling to provide diagnostic information in large scale assessment*. Paper presented at the annual meeting of the national council of measurement in education, San Diego, CA, April 2009.

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

Beguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional irt models. *Psychometrika*, *66*, 471–488.

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 of measurement in education, San Fransisco, CA, April 2006.

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. (2010, in press). *Subscores based on multidimensional item response theory* (ETS Research Rep.). Princeton, NJ: ETS.

Haberman, S.J., Sinharay, S., & Puhan, G. (2008). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95.

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.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Luecht, R.M., Gierl, M.J., Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual meeting of the national council on measurement in education, San Francisco, CA, April 2006.

Puhan, G., Sinharay, S., Haberman, S.J., & Larkin, K. (2010, in press). The utility of augmented subscores in a licensure exam: an evaluation of methods using empirical data. *Applied Measurement in Education*.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reckase, M.D. (2007). Multidimensional item response theory. In Rao, C.R., & Sinharay, S. (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–642). Amsterdam: North-Holland.

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

Sinharay, S. (2010, in press). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*.

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*, 21–28.

Thissen, D., Nelson, L., & Swygert, K.A. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—approximation methods for scale scores. In Thissen, D., & Wainer, H. (Eds.), *Test scoring* (pp. 293–341). Hillsdale: Lawrence Erlbaum.

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 Thissen, D., & Wainer, H. (Eds.), *Test scoring* (pp. 343–387). Hillsdale: 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, April 1987.

Manuscript Received: 13 JAN 2009
Final Version Received: 5 NOV 2009
Published Online Date: 27 MAR 2010