*Article*

# A Comparison of Four Methods of IRT Subscoring

Applied Psychological Measurement
35(4) 296–316

DOI: 10.1177/0146621610378653
http://apm.sagepub.com

**Jimmy de la Torre[1], Hao Song[2], and Yuan Hong[3]**

## Abstract

Lack of sufficient reliability is the primary impediment for generating and reporting subtest scores. Several current methods of subscore estimation do so either by incorporating the correlational structure among the subtest abilities or by using the examinee's performance on the overall test. This article conducted a systematic comparison of four subscoring methods—multidimensional scoring (MS), augmented scoring (AS), higher order item response model scoring (HO), and objective performance index scoring (OPI)—by examining how test length, number of subtests or domains, and correlation between the abilities affect the subtest ability estimation. The correlation-based methods (i.e., MS, AS, and HO) provided largely similar results, and performed best under conditions involving multiple short subtests and highly correlated abilities. In most of the conditions considered, the OPI method performed poorer compared to other methods on both ability estimates and proportion correct scores. Real data analysis further underscores the similarities and differences between the four subscoring methods.

## Keywords

item response theory, multidimensional IRT, higher order IRT, augmented scoring, objective performance index, ability estimation, Bayesian estimation, Markov chain Monte Carlo

Tests with multiple components are common in large-scale assessments. The components usually consist of subsets of items that measure specific content or process attributes beyond an overall ability. Although the overall ability estimate is useful for important decisions, the domain ability estimates complement the overall ability estimate by providing finer grained diagnosis of examinees' strengths and weaknesses. To make valid inferences about a student's attributes from the student's responses to items in the subtest domains, reliable subscores should be obtained. However, because of the small number of items within the subtest sections, lack of sufficient reliability is the primary impediment for generating and reporting subtest scores.

It is also not uncommon for educational assessments to involve a battery of tests that measure multiple correlated abilities. However, in practice, these tests are typically scored one test at

[1]Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[2]American Board of Internal Medicine, Philadelphia, PA, USA
[3]American Institutes for Research, USA

**Corresponding Author:**
Jimmy de la Torre, 10 Seminary Place, New Brunswick, NJ 08901, USA
Email: j.delatorre@rutgers.edu

a time, in that only responses specific to each test are used in estimating the examinees' abilities. Because the correlation structure of the abilities is relevant to examinees' standings on the tests, ignoring this information leads to suboptimal ability estimates.

A number of more efficient estimation procedures have been developed recently to improve the reliability and optimality of ability estimation. Several procedures capitalize on information found in the correlational structure of the abilities to improve item response theory (IRT) estimation. By incorporating the correlation structure in the estimation procedure using a multidimensional framework, de la Torre (2008), de la Torre and Patz (2005), and Wang, Chen, and Cheng (2004) obtained ability estimates from dichotomous and polytomous responses that are more accurate and precise compared to those obtained using one test at a time. The same framework when applied to item selection and scoring can provide substantial gains in the efficiency of computerized adaptive testing (Li & Schafer, 2005; Segall, 1996; Wang & Chen, 2004).

Wainer et al. (2001) provide an alternative way of using the correlational information to augment subscores. Their procedure relies on the test reliabilities and intertest correlations to estimate the correlations between the abilities, which in turn are used in computing the empirical Bayes estimates of the abilities. The method is a multivariate extension of Kelly's (1927) regressed scores and can be used in conjunction with a variety of score types: conventional summed score, scale score, and IRT score; the method has been used to augment subscores of the North Carolina End-of-Grade Mathematics Tests (Thissen & Edwards, 2005). When involving expected a posteriori (EAP; Bock & Aiken, 1981) estimates, this method and the multidimensional IRT scoring method above by de la Torre and Patz (2005) produce results that are practically indistinguishable from one another.

Still another procedure that accounts for the correlational structure of the abilities is a method based on the higher order IRT (HO-IRT; de la Torre & Song, 2009). The multidimensional method by de la Torre and Patz (2005) is identical to the HO-IRT method except for one important aspect: The former estimates an unconstrained correlational structure, whereas the latter constrains the correlations to have a one-factor structure using a higher order ability. Although more constrained in some respects, the HO-IRT model is more in line with the hierarchical ability structure well accepted in psychological research and practice (e.g., Carrol, 1993; Cronbach & Snow, 1977). In addition to improved subtest scores, the model allows for the overall ability to be simultaneously estimated within the same framework.

Instead of the correlations between the abilities, the method proposed by Yen (1987) uses the examinee's performance on the overall test to improve scores on the subsections of the test. Specifically, the overall ability estimate is first computed and used as ''prior information'' to improve estimation of the true score (proportion correct) in a particular test objective. This method results in objective scores (called the objective performance indexes [OPIs]) that have smaller standard errors and, thus, narrower credibility intervals. The OPI can be used with different types of responses (Yen, Sykes, Ito, & Julian, 1997) and has been used in the Comprehensive Tests of Basic Skills published by CTB/McGraw-Hill (CTB Technical Report, 1991).

A common thread runs through the different methods above, namely, the use of ancillary information to improve subscoring. Although ancillary or collateral information in IRT applications can pertain to additional information about the items (see Mislevy & Sheehan, 1989; Mislevy, Sheehan, & Wingersky, 1993; Tang & Eignor, 2001, for examples), the ancillary information involved above pertains to the examinees. Moreover, this type of examinee information can be classified as *in-test* ancillary information in that it is inherent in the test and can be found with the examinees' item responses. In these examples, which involved multiple subtests or domains, responses to tests in other domains are used as collateral information for the ability measured in a particular domain. In contrast, *out-of-test* ancillary information is a type of information that needs to be collected in addition to the item responses. Examples of this type of ancillary information are demographic

variables such as sex, age, and race, and educational variables such as grade level and courses taken. When available, de la Torre (2009) has shown that *out-of-test* ancillary information can improve ability estimation, particularly when combined with the *in-test* ancillary information.

Although the four methods—multidimensional scoring (MS), augmented scoring (AS), HO-IRT scoring (HO), and OPI scoring (OPI)—have been documented to improve subscoring, they have not been compared systematically with one another. Thus, the primary goal of this article is to examine how these methods compare with each other as factors affecting ability estimation (e.g., length of tests, number of tests, and correlational structure of the abilities) are varied. Within the multidimensional scoring framework, several procedures are available. But in this study, we chose to implement the multidimensional scoring method proposed by de la Torre and Patz (2005) because of its generality. Other methods of improving subscores are available (e.g., *within-item* multidimensional model, Reckase, 1996; bifactor model, Gibbons & Hedeker, 1992); however, these four methods were selected because of their flexibility (i.e., the methods can easily be modified in conjunction with different IRT models) and the ease by which the subtest abilities can be interpreted (i.e., only a single latent trait is involved in each subtest). Although subtests and tests in a battery differ in some respects (e.g., the abilities measured by the former are generally more highly correlated, whereas the latter tests are longer), the two types of test will be treated interchangeably for the purposes of this article. That is, although the article focuses on issues pertaining to subtest scores, the subscoring methods and results are equally applicable to battery scores.

The remaining sections of the article are laid out as follows: The next section provides an outline of how subscores are obtained using the four methods; the third section of the article uses simulated data to compare the quality of estimates of the four methods as a function of factors affecting ability estimation; the fourth section presents the results of a real data analysis using the four methods; finally, the last section presents a brief summary of the different findings, and discusses the practical implications of using these methods.

## Subscoring Methods

### Notation

Different notations have been used in conjunction with the multidimensional scoring, HO-IRT scoring, augmented scoring, and OPI scoring procedures. To facilitate the understanding and comparability of the different procedures, this article adopted the common notation described below.

As with other related works (e.g., de la Torre & Patz, 2005), we invoked the the simple structure assumption. In so doing, the multidimensional model by Reckase (1996) simplifies to the three-parameter logistic (3PL) model and is written for the purposes of this article as follows:

$$P_{j(d)}(\theta_{i(d)}) = P(X_{ij(d)} = 1|\theta_{i(d)}, \alpha_{j(d)}, \beta_{j(d)}, \gamma_{j(d)})$$

$$= \gamma_{j(d)} + (1 - \gamma_{j(d)}) \frac{\exp[1.7\alpha_{j(d)}(\theta_{i(d)} - \beta_{j(d)})]}{1 + \exp[1.7\alpha_{j(d)}(\theta_{i(d)} - \beta_{j(d)})]}, \quad (1)$$

where

$P(X_{ij(d)} = 1|\theta_{i(d)}, \alpha_{j(d)}, \beta_{j(d)}, \gamma_{j(d)})$ is the probability of examinee $i$ responding correctly to item $j$ of dimension $d$;

$\theta_{i(d)}$ is the $d$th component of the ability vector $\boldsymbol{\theta}_i$. (i.e., $\boldsymbol{\theta}_{i\cdot} = \{\theta_{i(d)}\}$);

$\alpha_{j(d)}$, $\beta_{j(d)}$, and $\gamma_{j(d)}$ are the discrimination, difficulty, and guessing parameters, respectively, of the $j$th item of dimension $d$;
$i = 1, \ldots, I$ (the total number of examinees);
$j = 1, \ldots, J$ (the total number of items);
$d = 1, \ldots, D$ (the number of dimensions);
$j(d) = 1(d), \ldots, J(d)$; and
$\sum_{d=1}^{D} J(d) = J$.

With the simple structure assumption, the item response $X_{ij(d)}$ has the likelihood

$$L_{j(d)}(\theta_{i(d)}) = (P(X_{ij(d)} = 1|\theta_{i(d)}, \alpha_{j(d)}, \beta_{j(d)}, \gamma_{j(d)}))^{X_{ij(d)}} (1 - P(X_{ij(d)} = 1|\theta_{i(d)}, \alpha_{j(d)}, \beta_{j(d)}, \gamma_{j(d)}))^{1-X_{ij(d)}}. \quad (2)$$

Let $\boldsymbol{X}_i$ represent the response vector of examinee $i$ across the $J$ items. By assuming conditional independence of the responses given the domain abilities, the corresponding likelihood of this vector is

$$L_i(\boldsymbol{X}_i|\boldsymbol{\theta}_{i\cdot}) = \prod_{d=1}^{D} \prod_{j(d)=1}^{J(d)} L_{j(d)}(\theta_{i(d)}). \quad (3)$$

Finally, the likelihood of the data matrix $\boldsymbol{X}$ given $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{i\cdot}\}$, the $N \times D$ matrix of domain-specific abilities, is

$$L(\boldsymbol{X}|\boldsymbol{\Theta}) = \prod_{i=1}^{I} L_i(\boldsymbol{X}_i|\boldsymbol{\theta}_{i\cdot}) = \prod_{i=1}^{I} \prod_{d=1}^{D} \prod_{j(d)=1}^{J(d)} L_{j(d)}(\theta_{i(d)}). \quad (4)$$

The steps in obtaining HO, MS, AS and OPI subscores are outlined below. Additional details concerning these procedures can be found in de la Torre and Song (2009), de la Torre and Patz (2005), Wainer et al. (2001), and Yen (1987), respectively.

## HO-IRT Scoring

The HO-IRT approach employs a hierarchical Bayesian framework and has the following model formulation:

$$\theta_i \sim N(0, 1) \quad (5)$$

$$\lambda_d \sim U(-1, 1) \quad (6)$$

$$\theta_{i(d)}|\theta_i, \lambda_d \sim N(\lambda_d \theta_i, 1 - \lambda_d^2). \quad (7)$$

where $\theta_i$ is the overall ability of examinee $i$, and $\lambda_d$ is the coefficient in regressing $\theta_{i(d)}$ on $\theta_i$. Let $\boldsymbol{\theta} = \{\theta_i\}$ and $\boldsymbol{\lambda} = \{\lambda_d\}$ be the $N \times 1$ vector of higher order abilities, and the $D \times 1$ vector of regression parameters, respectively. The joint posterior distribution of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\Theta}$ under the HO-IRT approach is

$$P(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}|\boldsymbol{X}) = L(\boldsymbol{X}|\boldsymbol{\Theta}) P(\boldsymbol{\Theta}|\boldsymbol{\theta}, \boldsymbol{\lambda}) P(\boldsymbol{\theta}) P(\boldsymbol{\lambda}). \quad (8)$$

Samples from Equation 8 can be obtained using Markov chain Monte Carlo (MCMC) simulation and are used in obtaining the parameter estimates $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\lambda}}$, and $\tilde{\boldsymbol{\Theta}}^{(\mathrm{HO})}$.

### Multidimensional Scoring

The hierarchical Bayesian formulation of the multidimensional scoring approach is as follows:

$$\boldsymbol{\Sigma} \sim Inv - Wishart_{v_0}(\boldsymbol{\Lambda}_0^{-1}) \tag{9}$$

$$\boldsymbol{\theta}_{i\cdot} | \boldsymbol{\Sigma} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}). \tag{10}$$

The joint posterior distribution of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ under the multidimensional scoring procedure is

$$P(\boldsymbol{\Sigma}, \boldsymbol{\Theta}|\boldsymbol{X}) = L(\boldsymbol{X}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\boldsymbol{\Sigma})P(\boldsymbol{\Sigma}). \tag{11}$$

As in the HO-IRT approach, the parameter estimates $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Theta}}^{(\mathrm{MS})}$ can also be obtained using MCMC.

### Augmented Scoring

The following steps outline how IRT ability estimates can be augmented.

Step 1. Obtain the conventional EAP estimate $\tilde{\theta}_{i(d)}$.

Step 2. Transform the estimate into the unregressed estimate $\tilde{\theta}^*_{i(d)}$ as follows:

$$\tilde{\theta}^*_{i(d)} = \frac{\tilde{\theta}_{i(d)}}{\rho_d}, \tag{12}$$

with $\rho_d$ representing the reliability of subtest $d$. The subtest reliability is computed as

$$\rho_d = \frac{V(\tilde{\boldsymbol{\theta}}_d)}{V(\tilde{\boldsymbol{\theta}}_d) + \overline{V(\tilde{\boldsymbol{\theta}}_d|\boldsymbol{X}_d)}}, \tag{13}$$

where $V(\tilde{\boldsymbol{\theta}}_d)$ and $\overline{V(\tilde{\boldsymbol{\theta}}_d|\mathbf{X}_d)}$ are the variance of EAP estimates and average posterior variance of the ability in subtest $d$, respectively.

Step 3. Compute $\mathbf{S}^u$, the covariance matrix of unregressed ability estimates across the $D$ domains.

Step 4. Define $\mathbf{S}^c = \mathbf{S}^u - \mathbf{D}$ to be the covariance matrix corrected for reliability, where $D$ is a diagonal matrix whose $d$th nonzero entry is $(1 - \rho_d)s^u_{dd}$.

Step 5. The augmented estimate of examinee $i$'s ability vector is given by

$$\tilde{\boldsymbol{\theta}}^{(\mathrm{AS})}_{i\cdot} = \overline{\tilde{\boldsymbol{\theta}}^*} + \mathbf{S}^c(\mathbf{S}^u)^{-1}(\tilde{\boldsymbol{\theta}}^*_{i\cdot} - \overline{\tilde{\boldsymbol{\theta}}^*}), \tag{14}$$

where $\overline{\tilde{\boldsymbol{\theta}}^*}$ is the mean vector of the unregressed ability estimates.

### OPI Scoring

The OPI score of examinee $i$ on subtest $d$ applied to the EAP can be obtained as follows:

Step 1. Derive the conventional EAP estimate of the overall ability $\tilde{\theta}_i$ using the responses to the $J$ items.

Step 2: Find $\tilde{\pi}_{i(d)} = \sum_{j(d)=1}^{J(d)} P_{j(d)}(\tilde{\theta}_{j(d)})/J(d)$, the estimated expected proportion of correct responses in subtest $d$.

Step 3: Approximate $\mu(\tilde{\pi}_{i(d)}|\theta_i)$ by $\tilde{\pi}_{i(d)}$, and $\sigma^2(\tilde{\pi}_{i(d)}|\theta_i)$ by

$$\left[\sum_{j(d)=1}^{J(d)} P'_{j(d)}(\tilde{\theta}_i)/J(d)\right]^2 \Big/ \sum_{d=1}^{D}\sum_{j(d)=1}^{J(d)} \frac{[P'_{j(d)}(\tilde{\theta}_i)]^2}{P_{j(d)}(\tilde{\theta}_i)[1-P_{j(d)}(\tilde{\theta}_i)]},$$

where $P'_{j(d)}(\theta)$ is the derivative of $P_{j(d)}(\theta)$ with respect to $\theta$.

Step 4: Find $J^*_{i(d)} = \mu(\tilde{\pi}_{i(d)}|\theta_i)[1-\mu(\tilde{\pi}_{i(d)}|\theta_i)]/\sigma^2(\tilde{\pi}_{i(d)}|\theta_i) - 1$.

Step 5: Compute

$$Q_i = \sum_{d=1}^{D} \frac{[T_{i(d)}/J(d) - \tilde{\pi}_{i(d)}]^2/J(d)}{\tilde{\pi}_{i(d)}(1-\tilde{\pi}_{i(d)})},$$

where $T_{i(d)} = \sum_{j(d)=1}^{J(d)} X_{ij(d)}$, and set $J^*_{i(d)} = 0$ if $Q_i > \chi^2(D, \alpha^*)$, the $(1-\alpha^*)$ quantile of the $\chi^2$ distribution with $D$ degrees of freedom.

Step 6: Given the response vector $\boldsymbol{X}_{i(d)}$, the posterior distribution of $\tilde{\pi}_{i(d)}$ is assumed to be beta-distributed with the parameters $\nu_{i(d)} = \tilde{\pi}_{i(d)}J^*_{i(d)} + T_{i(d)}$ and $\upsilon_{i(d)} = [1-\tilde{\pi}_{i(d)}]\ J^*_{i(d)} + J(d) - T_{i(d)}$. It follows that the posterior mean of $\tilde{\pi}_{i(d)}$ is

$$\mu(\tilde{\pi}_{i(d)}|\boldsymbol{X}_{i(d)}) = \frac{\nu_{i(d)}}{\nu_{i(d)}+\upsilon_{i(d)}} = \frac{\tilde{\pi}_{i(d)}J^*_{i(d)} + T_{i(d)}}{J^*_{i(d)} + J(d)}.$$

Step 7: Because $\pi_{i(d)}$ and $\theta_{i(d)}$ are monotonically related, the former can be obtained from the latter. Specifically, the OPI-based subtest ability is estimated as

$$\tilde{\theta}^{(OPI)}_{i(d)} = Inverse[\mu(\tilde{\pi}_{i(d)}|\boldsymbol{X}_{i(d)})].$$

Although other parameter estimates are also available, only $\tilde{\theta}^{(HO)}_{i(d)}$, $\tilde{\theta}^{(MS)}_{i(d)}$, $\tilde{\theta}^{(AS)}_{i(d)}$, and $\tilde{\theta}^{(OPI)}_{i(d)}$ from these methods were of interest in this study. Finally, in closing this section, we note the following about the four methods: (1) The OPI was originally designed for estimating the subtest true score using maximum likelihood estimation, but the use of EAP and the conversion of $\pi_{i(d)}$ to $\theta_{i(d)}$ were necessary for comparison purposes; (2) although the overall scores in HO and OPI procedures may carry the same interpretation, they are qualitatively different—the former is a higher order entity based on lower order ability estimates, whereas the latter is estimated directly from the item responses; and (3) because the HO and MS procedures require MCMC to be implemented, they are computationally more demanding than the AS and OPI procedures.

## Simulation Study

The number of examinees was found to have little effect on ability estimates for this type of estimation procedure (de la Torre & Patz, 2005; de la Torre & Song, 2009). For this reason, a fixed sample size, $N = 1000$, was used in the simulation study. Three factors were then considered: (a) number of subtests or domains ($D = 2$ and 5), (b) number of items in a domain ($J = 10$, 20, and 30), and (c) correlation between the abilities ($\rho = 0.0$, 0.4, 0.7, and 0.9). Fully crossing the levels of these factors yielded 24 conditions.

**Table 1.** Simulation Study Item Parameters

| Item | α | β | γ |
|---|---|---|---|
| 1 | 0.90 | 0.95 | 0.18 |
| 2 | 0.50 | 0.13 | 0.18 |
| 3 | 1.22 | 0.21 | 0.27 |
| 4 | 1.13 | –0.75 | 0.06 |
| 5 | 0.69 | 0.34 | 0.13 |
| 6 | 0.79 | –1.60 | 0.20 |
| 7 | 1.24 | 1.17 | 0.12 |
| 8 | 0.52 | –1.78 | 0.00 |
| 9 | 1.01 | 0.95 | 0.21 |
| 10 | 1.12 | –0.05 | 0.25 |

In terms of complexity (i.e., number of structural and incidental parameters), the HO-IRT is the most complex; thus, it was used in generating the simulated data. The item parameters in this study were obtained from a pool of 550 national standardized mathematics items. Ten items with mean information function closest to the mean information function of the entire item pool were selected. (See Table 1 for the item parameters.) To keep the mean test information function constant, these 10 items were replicated to produce longer tests of twenty or more items. For more straightforward comparisons, the same items were used for all the subtests. Finally, the item parameters were fixed in this study.

To generate the item responses using the HO-IRT model, the overall ability $\theta_i$ was first generated from $N(0, 1)$. Given the overall ability and correlation between subtest abilities $\rho$, the $d$th domain ability $\theta_{i(d)}$ was then generated from $N(\lambda\theta_i, 1 - \lambda^2)$, where the regression coefficient $\lambda = \sqrt{\rho}$. The regression coefficient $\lambda$ did not require the subscript $d$ because of the compound symmetry assumption invoked in this study. Finally, given $\theta_{i(d)}$ and the item parameters, responses to the $J$ items in subtest $d$ were simulated. Because the study focused on ability estimation, and because ability replicates naturally existed by design (i.e., $N = 1000$), only a single data set was generated and analyzed for each condition.

To summarize the characteristics of the generated data, the following statistics were computed: $p_{j(d)}$ (observed proportion correct), $Corr(j(d), T_{-j(d)})$ (item–rest score correlation), and $Corr(T_d, T_{d'})$ (subtest total correlation). The computation was taken across $D = 2$ and $D = 5$ because, in addition to greater stability, the number of dimensions was not expected to have any bearing on these statistics. The results in Table 2 indicate that the items were of moderate difficulty—the mean $p_{j(d)}$ is about 0.58. Moreover, the observed proportion correct was not affected by $J$ or $\rho$. The correlations between the item and rest score indicate that the items, on the average, were highly discriminating, and the item discriminations were only slightly higher for longer tests. Finally, the results also indicate that for $\rho > 0$, the correlation between subtest totals are higher when $\rho$ or $J$ was larger.

For the HO and MS methods, which obtained estimates via MCMC, we used chains that had the same number of burn-in, 5,000 draws, but different chain lengths. The chain lengths were determined to ensure that the structural parameters (i.e., the latent regression coefficients, correlation structure) had converged, here specifically to ensure that multivariate potential scale reduction factors (Brooks & Gelman, 1998) are less than 1.20. The domain ability estimates were based on the posterior means of the draws after the burn-in. The AS and OPI domain estimates were based on the procedures outlined in the previous section. In this study, we followed the critical value used by Yen (1987) and set $\alpha^* = 0.10$. The HO, MS, and AS domain ability estimates were converted into estimated proportion correct $\tilde{\pi}_{i(d)}$, whereas the OPI proportion

**Table 2.** Mean Observed Proportion Correct, Item–Rest Score Correlation, and Subtest Total Correlation

| | | ρ | | | |
|---|---|---|---|---|---|
| Statistic | $J$ | 0.0 | 0.4 | 7.0 | 0.9 |
| $p_{j(d)}$ | 10 | 0.58 | 0.58 | 0.58 | 0.58 |
| | 20 | 0.58 | 0.58 | 0.57 | 0.58 |
| | 30 | 0.58 | 0.58 | 0.58 | 0.57 |
| $Corr(j(d), T_{-j(d)})$ | 10 | 0.32 | 0.31 | 0.31 | 0.31 |
| | 20 | 0.35 | 0.34 | 0.34 | 0.35 |
| | 30 | 0.36 | 0.36 | 0.36 | 0.35 |
| $Corr(T_{(d)}, T_{(d')})$ | 10 | 0.00 | 0.26 | 0.47 | 0.58 |
| | 20 | 0.00 | 0.30 | 0.54 | 0.71 |
| | 30 | –0.02 | 0.34 | 0.57 | 0.76 |

correct estimates were converted to domain ability estimates $\tilde{\theta}_{i(d)}$. For the OPI conversion, the scale of the ability was truncated to be between –3 and 3. The quality of the domain estimates were measured by computing the correlation between the true and estimated abilities, the root mean squared error (RMSE) of the estimates across the examinees, and conditional bias and conditional mean absolute deviation (MAD) of $\tilde{\theta}_{i(d)}$. For the domain ability $\theta_{i(d)}$, the correlation and RMSE of the estimate were computed as $\mathrm{Cor}_{\forall i}(\theta_{i(d)}, \tilde{\theta}_{i(d)})$ and $\sqrt{\sum_{\forall i}(\tilde{\theta}_{i(d)} - \theta_{i(d)})^2}$; the conditional bias and conditional MAD were obtained for 12 intervals, namely, $(-\infty, -2.5)$, $[-2.5, -2), \ldots, [2.5, \infty)$, with the former computed as $\sum_{\forall \theta_{i(d)} \in (a,b)}(\tilde{\theta}_{i(d)} - \theta_{i(d)}) / I_{\forall \theta_{i(d)} \in (a,b)}$ and the latter $\sum_{\forall \theta_{i(d)} \in (a,b)} |\tilde{\theta}_{i(d)} - \theta_{i(d)}| / I_{\forall \theta_{i(d)} \in (a,b)}$, where $I_{\forall \theta_{i(d)} \in (a,b)}$ is the number of individuals in the interval $(a, b)$. The same statistics were obtained for $\pi_{i(d)}$. All computations and analyses in this article were run using Ox, an object-oriented matrix programming language that can be downloaded for free (Doornik, 2003). The code used in this study can be made available by contacting the first author.

## Results

### Ability Estimates

*Mean and SD of true and estimated abilities.* The results presented in this section were based on a single dimension. Tables 3 and 4 give the mean and *SD* of $\theta_{i(d)}$ and its various estimators, including the *uncorrected* EAP estimate of the ability. Mean estimates using EAP, HO, MS, and AS were for the most part similar to the mean of the true ability, and this discrepancy became less apparent when more, particularly shorter, tests were involved. In comparison, the means of the OPI estimates were typically smaller than the mean of the true ability across the different conditions.

In using EAP, shrinkage in estimates was observed, thus resulting in EAP estimates with smaller *SD* than the true ability. The relative size of the shrinkage was more apparent for shorter tests, but, as expected, not related to the number of and correlations between dimensions. Methods that incorporate the correlational information among the tests (i.e., HO, MS, and AS) produced estimates with smaller shrinkage, and their impact was more apparent for shorter but highly correlated tests. For example, when $D = 5$, $J = 10$, and $\rho = 0.9$, the EAP estimate had an *SD* of 0.83; in comparison, the HO, MS, and AS estimates had an *SD* of at least 0.91, which was closer to the *SD* of $\theta_{i(d)}$. In contrast, the OPI produced estimates that exhibited a different

**Table 3.** Mean of $\theta_{i(d)}$ and $\tilde{\theta}_{i(d)}$

| | | | | $\tilde{\theta}_{i(d)}$ | | | | |
|---|---|---|---|---|---|---|---|---|
| *D* | *J* | ρ | $\theta_{i(d)}$ | EAP | HO | MS | AS | OPI |
| 2 | 10 | 0.0 | 0.03 | 0.03 | 0.03 | 0.03 | 0.04 | –0.01 |
| | | 0.4 | 0.04 | 0.02 | 0.03 | 0.03 | 0.03 | 0.01 |
| | | 0.7 | –0.01 | –0.03 | –0.04 | –0.03 | –0.04 | –0.07 |
| | | 0.9 | 0.02 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 |
| | 20 | 0.0 | –0.00 | –0.03 | –0.03 | –0.03 | –0.04 | –0.05 |
| | | 0.4 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | | 0.7 | –0.01 | –0.02 | –0.03 | –0.03 | –0.03 | –0.05 |
| | | 0.9 | 0.02 | 0.03 | 0.03 | 0.03 | 0.04 | 0.02 |
| | 30 | 0.0 | –0.05 | –0.05 | –0.05 | –0.05 | –0.05 | –0.07 |
| | | 0.4 | 0.01 | 0.03 | 0.03 | 0.03 | 0.03 | 0.02 |
| | | 0.7 | 0.02 | 0.01 | 0.01 | 0.01 | 0.02 | 0.00 |
| | | 0.9 | 0.04 | –0.04 | –0.04 | –0.04 | –0.05 | –0.06 |
| 5 | 10 | 0.0 | –0.01 | 0.01 | 0.01 | 0.02 | 0.02 | –0.05 |
| | | 0.4 | –0.01 | 0.01 | 0.01 | 0.01 | 0.02 | –0.02 |
| | | 0.7 | 0.02 | –0.01 | 0.00 | 0.00 | –0.01 | 0.00 |
| | | 0.9 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | –0.00 |
| | 20 | 0.0 | –0.03 | –0.03 | –0.03 | –0.03 | –0.04 | –0.04 |
| | | 0.4 | –0.02 | –0.02 | –0.01 | –0.01 | –0.02 | –0.04 |
| | | 0.7 | –0.01 | –0.01 | –0.01 | 0.00 | –0.01 | –0.03 |
| | | 0.9 | –0.01 | –0.01 | 0.00 | –0.01 | –0.02 | –0.03 |
| | 30 | 0.0 | –0.01 | –0.03 | –0.03 | –0.03 | –0.03 | –0.05 |
| | | 0.4 | –0.02 | –0.02 | –0.02 | –0.02 | –0.02 | –0.03 |
| | | 0.7 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | | 0.9 | –0.06 | –0.08 | –0.07 | –0.07 | –0.09 | –0.11 |

Note: EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

pattern—the *SD* of the OPI estimate was larger, and in some cases, much larger than the *SD* of $\theta_{i(d)}$. As a whole, Tables 3 and 4 suggest that the HO, MS, and AS estimates can be expected to be similar to each other, but relatively different from the OPI estimates.

*Correlation with true ability.* Table 5 gives the correlation results between the true and estimated ability for the HO, MS, AS, and OPI procedures across the different conditions. For comparison purposes, the conventional EAP (i.e., estimates based solely on the responses to test items in a specific domain) results are also presented in addition to the four methods. (Technically, all the estimates in this article are various forms of EAP estimates. However, when there is no confusion, the term EAP will be used to refer to the conventional EAP.) Because EAP is unaffected by the correlations between the abilities, the EAP results were averaged across the different values of ρ. For example, for $D = 2$ and $J = 10$, $\sum_{\forall\rho} Cor(\theta_{i(d)}, \theta_{i(d)}^{(EAP)})/4 = 0.82$. The tabulated values for the four subscoring methods represent the difference between the results from the four subscoring methods and the EAP results. A positive difference indicates an improvement over the EAP results, whereas a negative difference indicates a deterioration (e.g., lower correlation, higher RMSE) relative to the EAP estimates.

The table shows that the correlations between the true and EAP-estimated abilities were 0.82, 0.89, and 0.93 when the abilities were estimated using 10-, 20-, and 30-item tests, respectively. These results were in line with the expectation that better estimates can be obtained with longer

**Table 4.** Standard Deviation of $\theta_{i(d)}$ and $\tilde{\theta}_{i(d)}$

| | | | | $\tilde{\theta}_{i(d)}$ | | | | |
|---|---|---|---|---|---|---|---|---|
| *D* | *J* | ρ | $\theta_{i(d)}$ | EAP | HO | MS | AS | OPI |
| 2 | 10 | 0.0 | 1.03 | 0.83 | 0.83 | 0.85 | 0.83 | 1.12 |
| | | 0.4 | 1.00 | 0.82 | 0.82 | 0.82 | 0.83 | 1.07 |
| | | 0.7 | 0.99 | 0.81 | 0.84 | 0.82 | 0.84 | 1.05 |
| | | 0.9 | 1.00 | 0.83 | 0.87 | 0.87 | 0.87 | 1.03 |
| | 20 | 0.0 | 1.00 | 0.90 | 0.91 | 0.91 | 0.90 | 1.09 |
| | | 0.4 | 0.99 | 0.87 | 0.87 | 0.84 | 0.87 | 0.99 |
| | | 0.7 | 0.99 | 0.89 | 0.91 | 0.90 | 0.90 | 1.02 |
| | | 0.9 | 1.00 | 0.90 | 0.93 | 0.93 | 0.93 | 1.02 |
| | 30 | 0.0 | 1.04 | 0.97 | 0.97 | 0.99 | 0.97 | 1.12 |
| | | 0.4 | 0.98 | 0.90 | 0.90 | 0.88 | 0.90 | 0.99 |
| | | 0.7 | 1.02 | 0.96 | 0.97 | 0.98 | 0.97 | 1.05 |
| | | 0.9 | 0.98 | 0.90 | 0.92 | 0.91 | 0.93 | 0.99 |
| 5 | 10 | 0.0 | 0.99 | 0.81 | 0.81 | 0.79 | 0.81 | 1.22 |
| | | 0.4 | 0.99 | 0.81 | 0.83 | 0.81 | 0.83 | 1.20 |
| | | 0.7 | 1.00 | 0.81 | 0.88 | 0.86 | 0.87 | 1.12 |
| | | 0.9 | 0.99 | 0.83 | 0.91 | 0.94 | 0.93 | 1.11 |
| | 20 | 0.0 | 0.99 | 0.89 | 0.89 | 0.88 | 0.89 | 1.16 |
| | | 0.4 | 0.99 | 0.88 | 0.88 | 0.88 | 0.89 | 1.12 |
| | | 0.7 | 0.95 | 0.84 | 0.87 | 0.83 | 0.87 | 1.06 |
| | | 0.9 | 1.01 | 0.91 | 0.95 | 0.95 | 0.96 | 1.13 |
| | 30 | 0.0 | 0.99 | 0.92 | 0.92 | 0.92 | 0.92 | 1.13 |
| | | 0.4 | 1.05 | 0.94 | 0.94 | 0.95 | 0.94 | 1.13 |
| | | 0.7 | 0.97 | 0.90 | 0.91 | 0.90 | 0.91 | 1.08 |
| | | 0.9 | 0.99 | 0.92 | 0.93 | 0.94 | 0.95 | 1.09 |

Note: EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

tests, but the differential benefit of adding a fixed number of items diminishes with higher test reliability. (Using the same 10 items for the different test lengths in this study makes the reliability of a subtest the same as the test length.) The table also shows that the HO, MS, and AS methods, which are procedures that exploit the the correlational structure of the abilities, gave very highly similar results; the OPI method, on the other hand, provided a very different pattern of results. For the three correlation-based methods, negligible to no improvements ($\leq 0.01$) were observed when $D = 2$ and $\rho \leq 0.4$ regardless of the test length. When at least moderately short tests were involved (i.e., $J \geq 20$) for the same number of domains, a $\rho = 0.7$ provided only negligible improvements. The best result for $D = 2$ occurred when the abilities are unreliably estimated (i.e., $J = 10$) and $\rho = 0.9$. Under this condition, using the information in the correlational structure resulted in estimates that correlated 0.87 with the true abilities, which represented an improvement of 0.05. In general, the improvement over the EAP estimates using the three subscoring methods was greater when ρ was higher but $J$ was lower.

Although the same pattern can be observed with changes in ρ and $J$, relatively higher correlations for the three methods were obtained when the number of subtests was increased from two to five. Incorporating the correlational structure in the ability estimation produced moderately large to large improvements (0.05-0.10) when tests were short and $\rho \geq 0.7$, and small improvements (0.03-0.04) when tests were longer and $\rho = 0.9$. Under the optimal condition (i.e., $D = 5$, $J = 10$, and $\rho = 0.9$), the ability estimates obtained using the HO, MS, and AS methods were

**Table 5.** Correlation Between $\theta_{i(d)}$ and $\tilde{\theta}_{i(d)}$

| | | | | Difference | | | |
|---|---|---|---|---|---|---|---|
| *D* | *J* | ρ | EAP | HO | MS | AS | OPI |
| 2 | 10 | 0.0 | 0.82 | 0.00 | 0.00 | 0.00 | –0.07 |
| | | 0.4 | | 0.01 | 0.01 | 0.01 | –0.05 |
| | | 0.7 | | 0.02 | 0.02 | 0.02 | –0.01 |
| | | 0.9 | | 0.05 | 0.05 | 0.05 | 0.02 |
| | 20 | 0.0 | 0.89 | 0.00 | 0.00 | 0.00 | –0.04 |
| | | 0.4 | | 0.00 | 0.00 | 0.00 | –0.03 |
| | | 0.7 | | 0.01 | 0.01 | 0.01 | –0.02 |
| | | 0.9 | | 0.03 | 0.03 | 0.03 | 0.01 |
| | 30 | 0.0 | 0.93 | 0.00 | 0.00 | 0.00 | –0.03 |
| | | 0.4 | | 0.00 | 0.00 | 0.00 | –0.03 |
| | | 0.7 | | 0.01 | 0.01 | 0.01 | –0.02 |
| | | 0.9 | | 0.02 | 0.02 | 0.02 | 0.00 |
| 5 | 10 | 0.0 | 0.82 | 0.00 | 0.00 | 0.00 | –0.10 |
| | | 0.4 | | 0.02 | 0.01 | 0.02 | –0.08 |
| | | 0.7 | | 0.05 | 0.05 | 0.05 | –0.03 |
| | | 0.9 | | 0.10 | 0.10 | 0.10 | 0.02 |
| | 20 | 0.0 | 0.89 | 0.00 | 0.00 | 0.00 | –0.03 |
| | | 0.4 | | 0.01 | 0.01 | 0.01 | –0.06 |
| | | 0.7 | | 0.02 | 0.02 | 0.02 | –0.04 |
| | | 0.9 | | 0.04 | 0.04 | 0.04 | –0.01 |
| | 30 | 0.0 | 0.93 | 0.00 | 0.00 | 0.00 | –0.02 |
| | | 0.4 | | 0.01 | 0.01 | 0.00 | –0.03 |
| | | 0.7 | | 0.01 | 0.01 | 0.01 | –0.02 |
| | | 0.9 | | 0.03 | 0.03 | 0.03 | –0.02 |

Note: Difference = difference between subscoring and EAP results; EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

only slightly inferior compared to the EAP estimates based on 30 items (i.e., 0.92 vs. 0.93). This finding gives one indication in practical terms of the degree of improvement that can be expected when using any of the correlation-based methods to estimate the abilities under the optimal condition.

Quite unexpectedly, the OPI methods resulted in poorer estimates in most of the conditions considered in this study. The results deteriorated when the correlation was less than 0.9, or when $D = 5$ and $J \geq 20$; when $D = 2$, $\rho = 0.9$ but $J \geq 20$, the improvements were marginal at best (i.e., $\leq 0.01$). The best results using OPI were obtained when the test was short and the correlation high, but these improvements were only marginal. Finally, it should be noted that the OPI ability estimates showed large deteriorations relative to the EAP estimates when short tests measuring uncorrelated abilities were involved.

*RMSE of* $\tilde{\theta}_{i(d)}$. For the most part, Table 6 shows the same pattern of results as the correlation table above. However, RMSE provides a different metric to compare the ability estimates of the four subscoring methods. When $D = 2$, the baseline RMSE for a 10-item test using EAP is 0.57; an additional 10 items resulted in a reduction of 0.12 in the RMSE, and a further reduction of 0.07 after adding 10 more items to the test. (The slightly different results for the $D = 5$ can be attributed to sampling variability.)

**Table 6.** Root Mean Squared Error of $\tilde{\theta}_{i(d)}$

| | | | | Difference | | | |
|---|---|---|---|---|---|---|---|
| *D* | *J* | ρ | EAP | HO | MS | AS | OPI |
| 2 | 10 | 0.0 | 0.57 | 0.00 | 0.00 | 0.00 | –0.17 |
| | | 0.4 | | 0.01 | 0.01 | 0.01 | –0.13 |
| | | 0.7 | | 0.04 | 0.04 | 0.03 | –0.06 |
| | | 0.9 | | 0.08 | 0.08 | 0.08 | 0.00 |
| | 20 | 0.0 | 0.45 | 0.00 | 0.00 | 0.00 | –0.12 |
| | | 0.4 | | 0.01 | 0.00 | 0.01 | –0.08 |
| | | 0.7 | | 0.02 | 0.02 | 0.02 | –0.06 |
| | | 0.9 | | 0.07 | 0.07 | 0.07 | 0.00 |
| | 30 | 0.0 | 0.38 | 0.00 | 0.00 | 0.00 | –0.11 |
| | | 0.4 | | 0.00 | 0.00 | 0.00 | –0.08 |
| | | 0.7 | | 0.02 | 0.02 | 0.02 | –0.07 |
| | | 0.9 | | 0.05 | 0.05 | 0.04 | 0.00 |
| 5 | 10 | 0.0 | 0.56 | 0.00 | 0.00 | 0.00 | –0.28 |
| | | 0.4 | | 0.02 | 0.02 | 0.02 | –0.24 |
| | | 0.7 | | 0.09 | 0.08 | 0.08 | –0.12 |
| | | 0.9 | | 0.18 | 0.17 | 0.18 | –0.04 |
| | 20 | 0.0 | 0.44 | 0.00 | 0.00 | 0.00 | –0.14 |
| | | 0.4 | | 0.01 | 0.01 | 0.01 | –0.17 |
| | | 0.7 | | 0.04 | 0.04 | 0.04 | –0.12 |
| | | 0.9 | | 0.11 | 0.11 | 0.10 | –0.07 |
| | 30 | 0.0 | 0.38 | 0.00 | 0.00 | 0.00 | –0.11 |
| | | 0.4 | | 0.02 | 0.02 | 0.01 | –0.10 |
| | | 0.7 | | 0.03 | 0.03 | 0.03 | –0.10 |
| | | 0.9 | | 0.09 | 0.09 | 0.09 | –0.09 |

Note: Difference = difference between subscoring and EAP results; EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

Again, the results for the three correlation-based methods were highly comparable, and the results can be summarized as follows: using any of the three subscoring methods, lower RMSE relative to EAP estimates were obtained when the test was shorter and the number of dimensions and the correlations between the dimensions were higher. Under the optimal condition (i.e., $J = 10$, $D = 5$, and $\rho = 0.9$), the three methods produced ability estimates that have the same RSME as the EAP estimates based on a 30-item test. In addition, when $D = 5$ and $\rho = 0.9$, and the test was not longer than 20 items, ability estimated using these methods reduced the RMSE by at least 0.10.

For all of the conditions considered in this study, the RMSEs of the OPI estimates were greater than or equal to their EAP counterparts. In fact, regardless of test length and correlation between abilities, the use of OPI resulted in deteriorations when $D = 5$, and the deteriorations were more evident and striking with shorter tests. When $D = 2$, the OPI estimates were comparable to the EAP estimates only when $\rho = 0.9$, whereas deteriorations were observed when the correlation was less than 0.9. The poor performance of the OPI method may be attributable to the fact that it was not designed to estimate $\theta_{i(d)}$, but $\pi_{i(d)}$. A separate section examines how the different subscoring methods perform in estimating the proportion of correct response.

*Conditional bias and MAD.* Figures 1 and 2 show the plots of the conditional bias and conditional MAD of $\tilde{\theta}_{i(d)}$ under the optimal condition (i.e., $D = 5, J = 10, \rho = 0.9$), whereas Figures

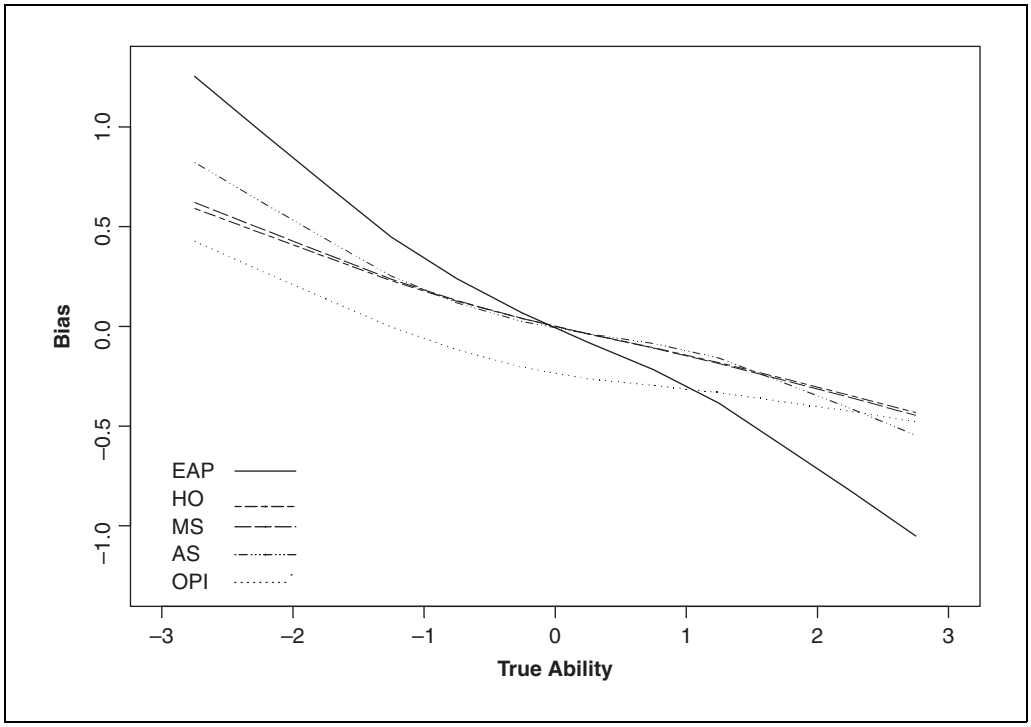


**Figure 1.** Conditional bias under the optimal condition ($D = 5, J = 10, \rho = 0.9$)
Note: EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

3 and 4 show the plots for the same statistics under the least optimal condition (i.e., $D = 2, J = 30, \rho = 0.0$). As shown in Figure 1, the conditional biases provided by the three correlation-based methods, for most of the continuum, were very similar to each other under the optimal condition; the exceptions can be observed outside $(-1.75, 1.75)$ where AS showed bias of slightly larger magnitude. The three methods provided conditional biases that were positive for negative $\theta_{i(d)}$ and negative for positive $\theta_{i(d)}$; the biases were of larger magnitude for more extreme values of $\theta_{i(d)}$. Compared with the EAP estimates, the HO, MS, and AS estimates were uniformly less biased, and the improvements were more evident for examinees with more extreme abilities. The conditional biases of the OPI estimates were markedly different from the rest of the methods. For one, it was more negatively biased than the rest—its bias was negative starting from about –1.25, whereas those of the remaining methods started from 0.00. For ability lower than –1.00 or so, the magnitude of the OPI bias was smallest compared to other methods; beyond this point, OPI had bias of larger magnitude for most part than those of the correlational methods but was better than EAP when $\theta_{i(d)} > 1.0$.

As with conditional bias, it is evident from Figure 2 that the conditional MADs of the correlational methods were uniformly better than that of the EAP. In addition, although the correlational methods had MADs that were largely similar, small discrepancies between these methods can be observed for more extreme abilities—HO had the smallest MAD followed by MS, then by AS. Compared to the correlational methods, OPI yielded the largest MAD for all but the extremely low-ability examinees; more importantly, OPI was worse than EAP for abilities in the middle range of the continuum where most of the examinees are typically located.

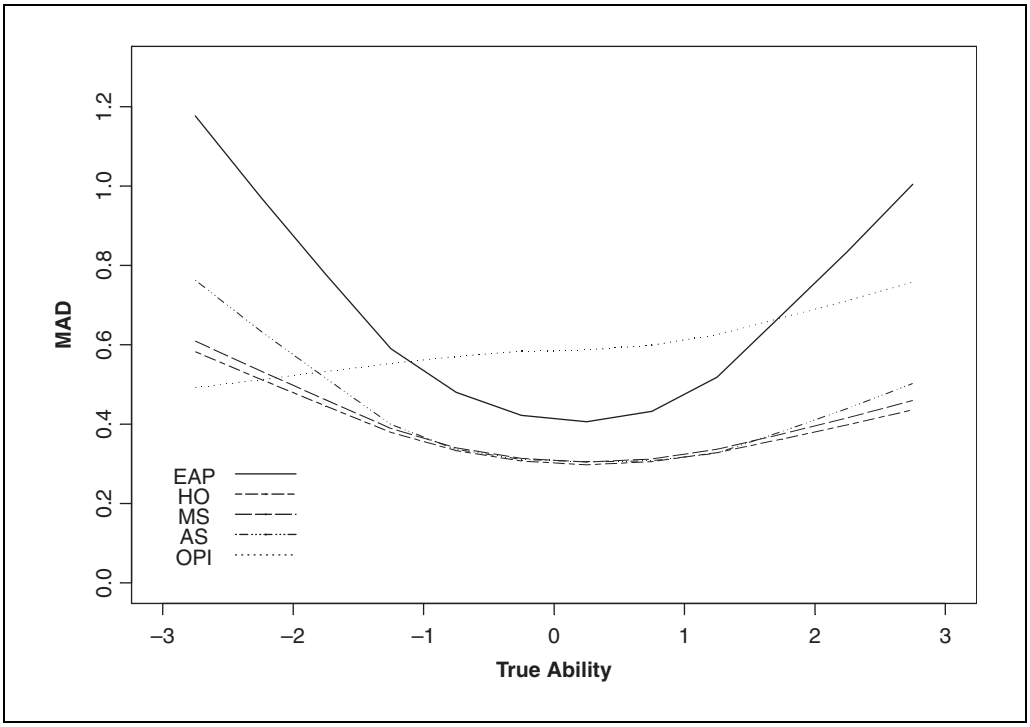


**Figure 2.** Conditional MAD under the optimal condition ($D = 5, J = 10, \rho = 0.9$)
Note: MAD = mean absolute deviation; EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

The conditional bias and conditional MAD under the least optimal condition (Figures 3 and 4) indicate that although all methods had the same pattern of bias and MAD (i.e., overestimation of negative abilities and underestimation of positive abilities, and larger discrepancies for more extreme abilities), the five methods exhibited several differences. First, EAP, HO, MS, and AS provided results that were similar to each other but not OPI. Second, despite the similarities, EAP, HO, MS, and AS results showed some divergence for extreme abilities. Specifically, results based on MS were slightly better. Last, although the magnitude of the bias obtained using OPI was uniformly the smallest, the corresponding MAD was uniformly the largest. This is an indication that the bias of the OPI estimates within an interval had greater variability (i.e., larger positive and negative biases).

*Estimated Expected Proportion Correct.* In contrast to the mean of the estimated abilities, the estimates of the expected proportion correct were highly similar to each other and to the true expected proportion correct. The mean $\pi_{i(d)}$ ranged from 0.56 to 0.58, and mean and maximum absolute differences of $\tilde{\pi}_{i(d)}$ across all methods and conditions were 0.00 and 0.01, respectively. The *SD* of $\tilde{\pi}_{i(d)}$ shows patterns similar to those in Table 4. That is, the EAP, HO, and MS estimates had smaller *SD*s, whereas the OPI, for the most part, had larger *SD*s relative to $SD(\pi_{i(d)})$. However, unlike $\tilde{\theta}_{i(d)}$, the differences in the *SD* of $\tilde{\pi}_{i(d)}$ were not as pronounced.

For the most part, the conclusions derived from $\theta_{i(d)}$ and $\tilde{\theta}_{i(d)}$ in Table 5 are applicable to the correlations between $\pi_{i(d)}$ and $\tilde{\pi}_{i(d)}$. For example, the EAP estimates (i.e., $\tilde{\theta}_{i(d)}$ converted to $\tilde{\pi}_{i(d)}$) had correlations of 0.83, 0.90, and 0.93 with the true proportion when the tests were 10, 20, and 30 items long, respectively—these are values very close to those obtained using $\tilde{\theta}_{i(d)}$.

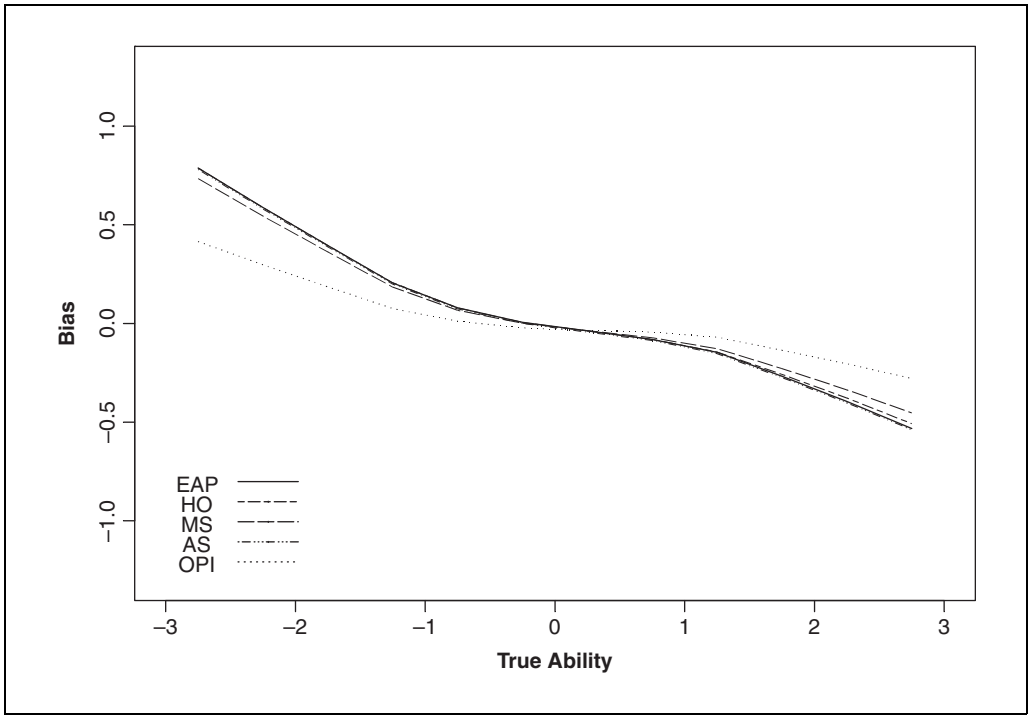


**Figure 3.** Conditional bias under the least optimal condition ($D = 2, J = 30, \rho = 0.0$)
Note: EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

Similarly, $RMSE(\tilde{\pi}_{i(d)})$ has the same pattern as $RMSE(\tilde{\theta}_{i(d)})$ but the magnitude for $\tilde{\theta}_{i(d)}$ was smaller because of the bounds in the probability values. For instance, the EAP-estimated proportion correct had an average RMSE of 0.10, 0.08, and 0.07 when they were based on 10-, 20-, and 30-item tests, respectively. Thus, the differences between the various methods were less pronounced when $\tilde{\pi}_{i(d)}$ was involved. The same effect as RSME can be observed for the conditional bias and conditional MAD of $\tilde{\pi}_{i(d)}$.

## Analysis of a Grade 9 Test Battery

We analyzed the responses of 2,255 examinees to a CTB/McGraw Hill Grade 9 test battery that has four content areas: Math (MA; 25 items), Math-Computation (MC; 20 items), Spelling (SP; 20 items), and Social Studies (SS; 25 items). These are the same data analyzed by de la Torre and Patz (2005) and de la Torre and Song (2009). The correlational structure of the abilities based on the multidimensional scoring analysis by de la Torre and Patz is given in Table 7. The correlation table shows that the highest and lowest correlations were between Math and Math-Computation (0.89), and between Math and Spelling (0.66), respectively, and the average correlation across the four content areas was about 0.75.

For this article, the abilities and the corresponding proportion correct for each examinee on the four content areas were estimated using the EAP and four subscoring methods. In the absence of the true ability and proportion correct, the different methods were compared using the

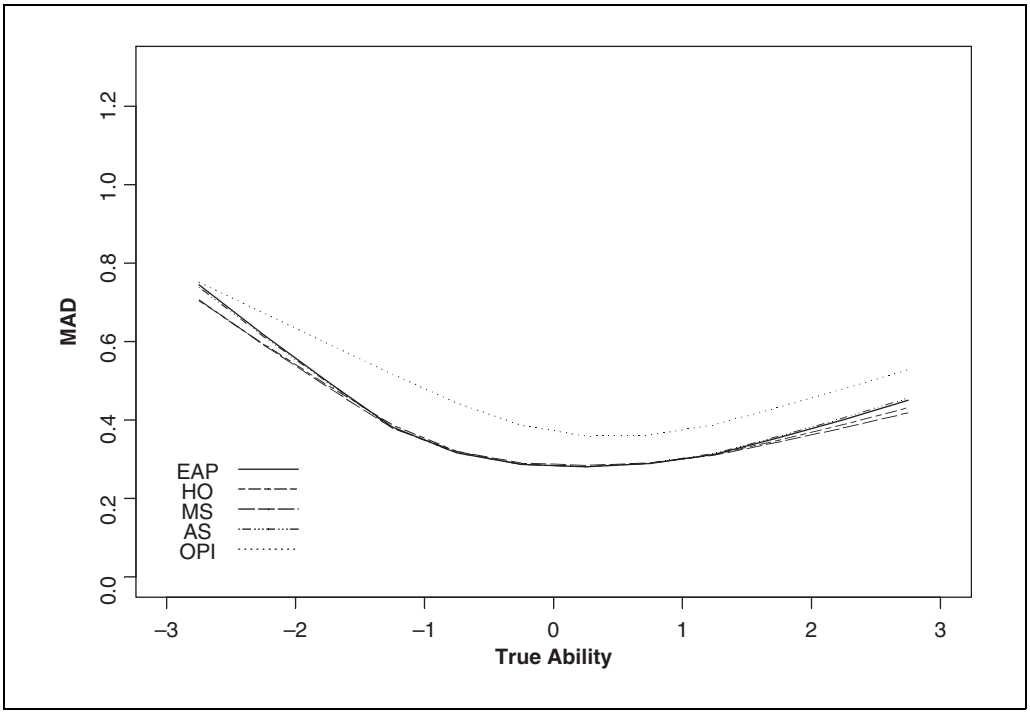


**Figure 4.** Conditional MAD under the least optimal condition ($D = 2, J = 30, \rho = 0.0$)
Note: MAD = mean absolute deviation; EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index.

characteristics of the distribution of the ability estimates. Specifically, summary statistics based on moments (*M* and *SD*) and quantiles (0.05, 0.50, and 0.95) were computed and compared.

## Results

The summary statistics for the ability estimates are given in Table 8. The mean ability estimates of the subscoring methods across the four content areas differed at most from the mean EAP estimates by 0.05 in absolute terms. The same can be said of the medians (i.e., 50th percentile)—the medians of the four subscoring methods were not very different from those of the EAP in that their maximum absolute difference was only 0.04. For both measures of central tendency, HO and MS showed a more similar pattern of over- and underestimation (relative to the EAP means) compared to the AS and OPI estimates.

The table also shows that all the subscoring methods produced estimates that were more variable than the EAP estimates across the four content areas. Of the four methods, the variabilities of the OPI estimates had the largest *SD* followed by the MS estimates, then by the HO estimates; AS had the most similar *SD* to the EAP. The same pattern can also be observed at the 5th percentile: The estimates using the subscoring methods were more extreme than the EAP estimates, and OPI had the most extreme 5th percentile estimates, followed by MS, then by HO, and finally by AS.

At the opposite end of the scale (i.e., 95th percentile) no clear pattern across the four subscoring methods was discernable. With the exception of SP, the EAP and subscoring method estimates were not very different at the 95th percentile—the absolute differences were less than or equal to 0.05. However, the differences between HO and EAP, and AS and EAP, were large

**Table 7.** Correlation Estimates for the Grade 9 Test Battery

| | MC | SP | SS |
|---|---|---|---|
| MA | 0.89 | 0.66 | 0.81 |
| MC | – | 0.70 | 0.76 |
| SP | – | – | 0.70 |

Note: MA = math; MC = math-computation; SP = spelling; SS = social studies.

**Table 8.** Summary Statistics for the Real Data $\tilde{\theta}_{i(d)}$

| | | | Content area | | | |
|---|---|---|---|---|---|---|
| Statistic | | Method | MA | MC | SP | SS |
| Moment | Mean | EAP | –0.02 | 0.02 | –0.08 | –0.16 |
| | | HO | –0.05 | –0.03 | –0.09 | –0.14 |
| | | MS | –0.06 | –0.03 | –0.10 | –0.15 |
| | | AS | –0.02 | 0.02 | –0.11 | –0.18 |
| | | OPI | –0.05 | –0.03 | –0.11 | –0.16 |
| | *SD* | EAP | 0.98 | 0.97 | 0.87 | 0.96 |
| | | HO | 1.02 | 1.02 | 0.92 | 0.99 |
| | | MS | 1.06 | 1.06 | 0.93 | 1.00 |
| | | AS | 1.00 | 0.99 | 0.90 | 0.98 |
| | | OPI | 1.10 | 1.12 | 1.10 | 1.07 |
| Quantile | 0.05 | EAP | –1.64 | –1.53 | –1.59 | –1.90 |
| | | HO | –1.80 | –1.78 | –1.74 | –1.94 |
| | | MS | –1.86 | –1.87 | –1.78 | –1.97 |
| | | AS | –1.68 | –1.57 | –1.72 | –1.93 |
| | | OPI | –1.98 | –1.96 | –1.98 | –2.06 |
| | 0.05 | EAP | 0.04 | 0.03 | –0.06 | –0.11 |
| | | HO | 0.02 | 0.03 | –0.05 | –0.09 |
| | | MS | 0.02 | 0.04 | –0.04 | –0.09 |
| | | AS | 0.03 | 0.06 | –0.07 | –0.13 |
| | | OPI | 0.04 | 0.02 | –0.02 | –0.07 |
| | 0.95 | EAP | 1.59 | 1.63 | 1.30 | 1.42 |
| | | HO | 1.54 | 1.59 | 1.40 | 1.41 |
| | | MS | 1.62 | 1.64 | 1.37 | 1.41 |
| | | AS | 1.59 | 1.65 | 1.34 | 1.37 |
| | | OPI | 1.58 | 1.63 | 1.55 | 1.46 |

Note: EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index; MA = math; MC = math computation; SP = spelling; SS = social studies.

and extremely large for the domain SP. To some extent, this can be attributed to the low correlation of SP with the other domains.

These results indicate that the ability estimates obtained using the four subscoring methods, on the average, were not too different from each other. However, they differed in variability and their estimation of low-ability examinees. In particular, the OPI showed a greater tendency of producing more extreme results in these respects. The OPI estimates can also be markedly different at the upper extreme, possibly when the domain does not correlate highly with other domains. To a large extent, these findings are consistent with those obtained in the simulation study.

**Table 9.** Summary Statistics for the Real Data $\tilde{\pi}_{i(d)}$

| Statistic | | Method | Content area | | | |
|---|---|---|---|---|---|---|
| | | | MA | MC | SP | SS |
| Moment | Mean | EAP | 0.58 | 0.57 | 0.54 | 0.64 |
| | | HO | 0.58 | 0.56 | 0.54 | 0.64 |
| | | MS | 0.58 | 0.56 | 0.53 | 0.64 |
| | | AS | 0.58 | 0.57 | 0.53 | 0.64 |
| | | OPI | 0.58 | 0.56 | 0.53 | 0.64 |
| | *SD* | EAP | 0.19 | 0.22 | 0.12 | 0.18 |
| | | HO | 0.20 | 0.22 | 0.13 | 0.19 |
| | | MS | 0.20 | 0.23 | 0.13 | 0.19 |
| | | AS | 0.20 | 0.22 | 0.12 | 0.19 |
| | | OPI | 0.20 | 0.23 | 0.15 | 0.20 |
| Quantile | 0.05 | EAP | 0.28 | 0.25 | 0.33 | 0.30 |
| | | HO | 0.26 | 0.22 | 0.32 | 0.29 |
| | | MS | 0.26 | 0.22 | 0.31 | 0.29 |
| | | AS | 0.27 | 0.24 | 0.32 | 0.30 |
| | | OPI | 0.25 | 0.21 | 0.30 | 0.28 |
| | 0.50 | EAP | 0.60 | 0.57 | 0.54 | 0.66 |
| | | HO | 0.59 | 0.57 | 0.54 | 0.66 |
| | | MS | 0.59 | 0.57 | 0.54 | 0.66 |
| | | AS | 0.60 | 0.57 | 0.53 | 0.66 |
| | | OPI | 0.60 | 0.56 | 0.54 | 0.67 |
| | 0.95 | EAP | 0.88 | 0.91 | 0.73 | 0.92 |
| | | HO | 0.88 | 0.91 | 0.75 | 0.91 |
| | | MS | 0.88 | 0.91 | 0.74 | 0.91 |
| | | AS | 0.88 | 0.91 | 0.74 | 0.91 |
| | | OPI | 0.88 | 0.91 | 0.77 | 0.92 |

Note: EAP = expected a posteriori; HO = higher order; MS = multidimensional scoring; AS = augmented scoring; OPI = objective performance index; MA = math; MC = math computation; SP = spelling; SS = social studies.

Similar to the results of the simulation study, although the patterns were still discernable, the different methods showed fewer discrepancies when compared in terms of the estimated proportion correct (see Table 9). Their mean and median estimates did not differ by more than 0.01 in absolute terms. The *SD*s of the correlation-based estimates were the same as or only slightly larger than the *SD*s of the EAP estimates. The OPI showed slightly larger *SD*s in the two mathematics areas, and noticeably larger *SD*s in the SP and SS areas.

For the examinees at the 5th percentile, the estimated proportions correct using the subscoring methods were less than or equal to the estimates using EAP (i.e., they were more extreme). MA and OPI had the largest discrepancies among the content areas and methods, respectively. At the 95th percentile, the four subscoring methods showed identical or almost identical estimates of $\tilde{\pi}_{i(d)}$ in the MA, MC, and SS content areas. Similar to $\tilde{\theta}_{i(d)}$, HO and OPI showed more noticeable discrepancies in SP.

## Summary and Discussion

This article systematically investigated the impact of three factors, namely, number of domains, test length, and correlation between domains, on four methods of IRT subscoring in estimating the

domain ability $\theta_{i(d)}$ and expected proportion correct $\pi_{i(d)}$. The simulation study showed that the correlation-based methods (i.e., HO, MS, and AS) gave highly comparable results across the different conditions except for extreme abilities where HO and MS may perform better. This finding is consistent with what de la Torre and Patz (2005) found in comparing MS and AS. The correlational-based methods provided better results relative to the conventional unidimensional EAP approach in situations that involved multiple short tests measuring highly correlated abilities. De la Torre and Song (2009), de la Torre and Patz (2005), and Edwards and Vevea (2006) obtained the same results in comparing EAP against HO, MS and AS, respectively. The improvements can be sizeable in comparing $\tilde{\theta}_{i(d)}$, but less so when comparing $\tilde{\pi}_{i(d)}$. In fact, estimates obtained using OPI were not necessarily better than their EAP counterparts. In some cases, OPI produced estimates that correlated much lower with the true ability and expected proportion correct, and had much larger RMSEs. Thus, among the methods considered in this study, including the conventional EAP, it can be argued that OPI produced estimates with the least desirable statistical qualities.

As noted earlier, for comparability purposes, our implementation of OPI in this study deviates from its traditional implementation in that, instead of MLE, we used the EAP estimates of the overall ability. However, our findings do not necessarily contradict previous findings. For one, the statistics used in this study focus on the posterior mean, whereas previous studies (e.g., Yen, 1987) have focused on the posterior variance. Because OPI estimates are based on the posterior mean and involve a proper prior, they cannot be unbiased (Gelman, Carlin, Stern, & Rubin, 2003; Lehmann & Casella, 1998). As the RMSE of the OPI estimates in this study has shown, the gain in precision using OPI is offset by the amount of bias introduced in the estimation process to the extent that the resulting estimates have poorer statistical qualities. For another, Monfils, Dawber, Han, and Henderson-Montero (2006) have also found that OPI estimates based on MLE of the overall ability have similar means but larger *SD*s compared to the AS estimates.

The three correlation-based methods in this study can be viewed as a more general approach in finding the conventional EAP estimate of the ability: Conventional EAP estimates can be obtained using these methods when the abilities are uncorrelated; substantially better estimates can be obtained when multiple short tests measuring highly correlated abilities are involved. This shows that although EAP estimates have desirable properties (i.e., smaller bias and standard error; Kim & Nicewander, 1993; Thissen & Orlando, 2001), further improvement can be achieved by taking into account the correlational structure of the abilities.

Except for extreme abilities where HO and MS may provide better estimates than AS, the choice between the three correlational-based methods may not be clear-cut if the decision is to be made solely on the basis of their statistical properties. Narrowing down the options requires that nonstatistical factors be considered. In terms of efficiency, AS requires the shortest time compared to HO and MS, which are MCMC based because of the relative complexity of the two models. It takes AS 3.43 seconds to obtain the ability estimates under the the condition $D = 2$, $J = 10$, and $\rho = 0.9$; with 10,000 iterations and a burn-in of 2,000 draws, the estimation times for MS and HO were approximately 40 and 42 times longer. Even with more efficient coding and fewer iterations, HO and MS can be expected to be 10 to 20 times computationally less efficient than AS. Although they are more time-consuming to implement, these procedures can be made more practicable in operational testing settings by using MCMC only for estimating the structural parameters; for the subsequent step of scoring examinees, possibly from different test administrations, more efficient maximization methods in conjunction with fixed structural parameters can be used. In addition, inefficiency notwithstanding, there are situations where MS and HO are more appropriate. For example, MS has a general framework that makes the model amenable to other sources of information. In a recent work, de la Torre (2009) shows that MS can be extended to incorporate out-of-test information. In some conditions, the resulting model produced better ability estimates than the current formulation of MS. For another, HO should be the

model of choice if a unified framework for obtaining the overall and domain ability estimates is of interest. Moreover, application to test battery calibration can easily be done using HO (see de la Torre & Hong, 2010). In general, identifying the most appropriate method of scoring, domain scoring or otherwise, requires bringing to bear statistical and pragmatic considerations, and perhaps, one's philosophical persuasion as well.

Finally, because of the ubiquity of test batteries or multiple subdomains within a test, methods that use in-test information can be useful in many practical testing situations. However, because estimates obtained using these methods have more complex interpretation in that they reflect abilities across multiple domains, their applications may not always be valid. One needs to be thoughtful about how the test scores will be used (Mislevy, 1987; Wainer et al, 2001). Specifically, this type of score is particularly well suited when a score profile is needed to determine a student's specific areas of strength and weakness. Scores from such a profile are more accurate and precise, and they can be used more reliably in determining the best learning trajectory for each student. But when the use of test scores has high-stakes implications (e.g., certification, competition), sufficiently reliable scores derived without the benefit of ancillary information would be more appropriate. However, as this dichotomy does not account for all possible testing scenarios, some situations (e.g., when students' overall performance and specific strengths and weakness need to be evaluated) may go beyond a single score type and require both the overall and domain scores to be reported for these scores to complement one another (de la Torre & Patz, 2005).

## Declaration of Conflicting Interests

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

## Funding

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

## References

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434-455.

Carrol, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitude and instructional methods*. New York, NY: Irvington.

CTB/McGraw-Hill. (1991). *Technical bulletin 1 of California Achievement Tests Forms C and D*. Monterey, CA: Author.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement, 32*, 355-370.

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement, 33*, 465-485.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267-285.

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

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*, 620-639.

Doornik, J. A. (2003). Object-oriented matrix programming using Ox (Version 3.1) [Computer software]. London, England: Timberlake Consultants Press.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*, 241-259.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman.

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

Kelly, T. L. (1927). *The interpretation of educational measurement*. New York: World Book.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587-599.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York, NY: Springer.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29*, 3-25.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81-91.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661-679.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55-78.

Monfils, L., Dawber, T., Han, N, & Henderson-Montero, D. (2006, April). *Supporting reform efforts through diagnostic subscore reports: Implications for schools*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Reckase, M. D. (1996). A linear logistic multidimensional model. In W. J. van der Linder & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Tang, K. L., & Eignor, D. R. (2001). *A study of the use of collateral statistical information in attempting to reduce TOEFL IRT item parameter estimation sample sizes* (TOEFL Technical Report). Princeton, NJ: Educational Testing Service.

Thissen, D., & Edwards, M. C. (2005, April). *Diagnostic scores augmented using multidimensional item response theory: Preliminary investigation of MCMC strategies*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Erlbaum.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2001). Augmented scores—''Borrowing strength'' to compute score based on small numbers of items. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 343-388). Mahwah, NJ: Erlbaum.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295-316.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

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

Yen, W. M., Sykes, R. C., Ito, K., & Julian, M. (1997, April). *A Bayesian/IRT index of objective performance for tests with mixed item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.