A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification

Lihua Yao and Keith A. Boughton, CTB/McGraw-Hill

Several approaches to reporting subscale scores can be found in the literature. This research explores a multidimensional compensatory dichotomous and polytomous item response theory modeling approach for subscale score proficiency estimation, leading toward a more diagnostic solution. It also develops and explores the recovery of a Markov chain Monte Carlo (MCMC) estimation approach to multidimensional item and ability parameter estimation, as well as subscale proficiency and classification rates. The simulation study presented here used real data-derived parameters from a large-scale statewide assessment with subscale score information under varying conditions of sample size and correlations between subscales (0.0, 0.1, 0.3, 0.5, 0.7, 0.9). It was found that to report accurate diagnostic information at the subscale level, the subscales need to be highly correlated, or a multidimensional approach should be implemented. MCMC methodology is still a nascent methodology in psychometrics; however, with the growing body of research, its future looks promising. *Index terms: multidimensional item response theory (MIRT); Bayesian estimation; MCMC; Domain score; OPI; subscale scores.*

The field of educational assessment has been dominated by unidimensional summative tests that report overall scores indicating students' levels of achievement in broadly defined content area domains, such as mathematics or language arts. Although there has been an increasing demand for assessments that directly assist students and teachers in understanding and responding to key strengths and weaknesses in specific content domains, this need for large-scale assessments that are more formative and informative in design remains largely unfulfilled.

Many testing programs (e.g., Scholastic Aptitude Test [SAT]; College Board) have introduced a diagnostic pretest, as in the PSAT, to assist students in determining which of the skills within a particular domain of knowledge needs improvement. CTB/McGraw-Hill's Tests of Adult Basic Education (TABE) is a diagnostic pretest for the General Education Diploma (GED) test in the United States. Operating under the assumption that a test can measure more than one trait or distinct cognitive dimension, numerous testing programs now require reporting of subscale scores for different objectives defined by the test design. For example, an overall mathematics score might be supplemented with scores for numbers and operations, algebra and functions, geometry and measures, and probability and statistics. Some programs report subscale scores on a simple number-correct score (or percentage correct), which in many cases is not adjusted for form difficulty and thus does not permit form-to-form or population-to-population comparisons. These simple scores also do not augment the subscores based on information from the other subscores and do not consider the relevance

Applied Psychological Measurement, Vol. 31 No. 2, March 2007, 1–23
DOI: 10.1177/0146621606291559

of departures from unidimensionality. A test designed solely for the purpose of scaling examinees on a single continuum may not be accurate or reliable at a more diagnostic level. To address this concern, several approaches have been introduced to help stabilize diagnostic subscale scores by augmenting data from the other subscales of the test.

Yen (1987) proposed an empirical Bayes procedure to reduce the error in subscale estimation by incorporating information from the total test score. The resulting objective performance index (OPI) has been widely used at CTB/McGraw-Hill. Wainer et al. (2001) proposed an augmentation to Yen's method, by allowing the information from the other subscale scores (they note that Yen only used the total test score) to help stabilize each diagnostic subscale score, with each item loading on only one subscale. Wainer et al. reported that this augmentation is well suited for tests that may be more multidimensional in nature, although the scores cannot be formerly equated at the subscale level.

Bock, Thissen, and Zimowski (1997) proposed an item response theory (IRT) estimation approach to subscale score estimation (i.e., domain score) that produced more accurate estimates than the percentage-correct score. Their domain score approach can be used for mastery classifications, and they note that although IRT scale scores are just as accurate, they do not lend themselves to the straightforward interpretation of the domain scores. They also note that this method can easily be used with multidimensional tests consisting of mixtures of dichotomous and polytomous items. This multidimensional extension is in fact much better suited for diagnostic subscale scoring and interpretation, allowing for more granularity; however, Bock et al. did not apply or conduct research on this multidimensional IRT (MIRT) extension.

Some other alternative scoring procedures for improving subscale score estimation, which are attempts to approximate the results that might be obtained through multidimensional modeling while still operating under a unidimensional paradigm, have been considered by a few researchers. Ackerman and Davey (1991) developed an approach that fits a unidimensional IRT model to multidimensional data and then attempts to improve trait estimation by using collateral information (i.e., information from secondary traits). Davey and Hirsh (1991) developed a procedure, which they termed the *concurrent calibration method*, that uses collateral parameters of items in the other subscales in the scoring of any particular subscale. Kahraman and Kamata (2004) combined Ackerman and Davey's (1991) and Davey and Hirsh's (1991) approaches to increase the precision of subscale scores, but they found that a high correlation between subscale scores is an essential requirement when trying to increase the subscale scores using out-of-scale information. If the out-of-scale items are highly discriminating, then a correlation as high as .9 would be needed.

MIRT is a growing methodology for modeling the relationships of examinees to a set of test items using a matrix of responses. Score estimation that taps into the relationships across several objectives, subscales, or test batteries (e.g., Wang, Chen, & Cheng, 2004) is more informative and reliable compared to a single scoring apparatus with simpler subscale scores, such as the number correct. Students, teachers, and parents would be able to tell which objective area students have mastered and in which area they can benefit from further instruction. Most important, with MIRT scaling procedures, objective scores are made comparable across forms and samples of examinees. Note that this linking is a critical aspect for new form development and has not yet been fully resolved for many of the diagnostic subscale modeling approaches.

There are two well-known MIRT estimation software programs for dichotomous data, one called TESTFACT (Wilson, Wood, & Gibbons, 1987) and the other NOHARM (Fraser, 1987). MIRT dichotomous models have been employed under a variety of conditions (Ackerman, 1994a, 1994b; Béguin & Glas, 2001; Reckase, 1985; Walker & Beretvas, 2003). However, many types of assessments currently contain a mixture of multiple-choice and constructed-response tasks that require different item response models (Lane, 2005). For example, Walker and Beretvas (2003)

demonstrated the usefulness of MIRT as a diagnostic tool, but given the lack of polytomous MIRT estimation software, they had to dichotomize all of the polytomous items and use NOHARM. MIRT parameterization software programs for tests with mixtures of dichotomous and polytomous items are relatively new and are (a) POLYFACT (Muraki, 1999), which uses marginal maximum likelihood (MML) estimation; (b) MicroFACT (Waller, 2002), which, like TESTFACT, employs factor analysis; and (c) a Bayesian multivariate item response theory (BMIRT; Yao, 2003) program, which adopts an Markov chain Monte Carlo (MCMC) solution.

Given the need for tests to be more diagnostic in purpose and product, as well as the fact that assessments are already too long, accurate subscale score estimation is of paramount importance. To this end, the present study introduces a confirmatory MIRT modeling approach that can be used in place of a simple number-correct or unidimensional IRT subscale scoring approach (e.g., OPI), based on simple structure, as defined by the content blueprint map. In addition, the approach developed here is not limited to multiple-choice (MC) tests but allows for a mixture of item response formats, including constructed-response (CR) items.

Method

This research examined several methods for determining examinees' objective scores by assessing their deviation from "true" objective scores under simulated conditions, with the generating parameters coming from real data. The accuracy of estimation for four approaches was assessed:

- Percentage correct on subscale number-correct scores (NC)
- Multidimensional IRT Bayesian subscale scores (BMIRTSS)
- Multidimensional IRT Bayesian domain subscale scores (BMIRTDS)
- OPI subscale scores

In addition, the BMIRT Bayesian estimation procedure was compared to an IRT pattern subscale scoring approach using maximum likelihood (MIRTPSS) estimation and a unidimensional IRT objective-level Bayesian scoring approach (UIRTOJSS), as it is important to thoroughly demonstrate the estimation accuracy in comparison with known scoring programs and procedures that have been used operationally.

Multidimensional Models

MIRT models can be characterized as either being compensatory or noncompensatory. The compensatory version allows the dimensions to interact, with a high ability on one dimension, compensating for a lower ability on a second dimension. However, within the noncompensatory version, one must be proficient in both abilities to obtain a higher score. The former is the focus of this research.

Suppose there are N examinees and J test items. The observable item response data are contained in a matrix $X = \{X_{ij}\}$, where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$. For convenience, let the responses of the ith examinee be a vector $\vec{X}_i = (X_{i1}, \ldots, X_{iJ})$. The ability parameters for examinees are

$$\theta = (\vec{\theta}_1, \ldots, \vec{\theta}_N)^T, \tag{1}$$

such that each $\vec{\theta}_i$ for $i = 1, 2, \ldots, N$ is a vector of dimension D, where D is the number of subscales or the number of dimensions hypothesized.

For a dichotomous item j, the probability of a correct response to item j for an examinee with ability $\vec{\theta}_i$ for the multidimensional three-parameter logistic (M-3PL; Reckase, 1997) model is

$$P_{ij1} = P(x_{ij} = 1 \mid \vec{\theta}_i, \vec{\beta}_j) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\vec{\beta}_{2j} \odot \vec{\theta}_i^T + \beta_{1j})}}, \tag{2}$$

where

- $x_{ij} = 0$ or 1 is the response of examinee i to item j.
- $\vec{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2jD})$ is a vector of dimension D of item discrimination parameters.
- β_{1j} is the scale difficulty parameter.
- β_{3j} is the scale guessing parameter.
- $\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$.

The parameters for the jth item are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j}). \tag{3}$$

For a polytomous item j, the probability of a response $k-1$ to item j for an examinee with ability $\vec{\theta}_i$ is given by the multidimensional version of the partial credit model (M-2PPC; Yao & Schwarz, in press):

$$P_{ijk} = P(x_{ij} = k-1 \mid \vec{\theta}_i, \vec{\beta}_j) = \frac{e^{(k-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{k} \beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{m} \beta_{\delta_t j})}}, \tag{4}$$

where

- $x_{ij} = 0, \ldots, K_j - 1$ is the response of examinee i to item j.
- $\vec{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2jD})$ is a vector of dimension D for the item discrimination parameters.
- $\beta_{\delta_k j}$ for $k = 1, 2, \ldots, K_j$ are the threshold or alpha parameters, $\beta_{\delta_1 j} = 0$, and K_j is the number of response categories for the jth item.

The parameters for the jth item are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta_2 j}, \ldots, \beta_{\delta_{K_j} j}). \tag{5}$$

Let

$$P_{ij} = P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j) = P_{ij1}^{1_{(X_{ij}=1)}} (1 - P_{ij1})^{1_{(X_{ij}=0)}}, \tag{6}$$

for a M-3PL item or

$$P_{ij} = P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j) = \prod_{k=1}^{K_j} P_{ijk}^{1_{(X_{ij}=k-1)}}, \tag{7}$$

for an M-2PPC item, where

$$1_{(X_{ij}=k)} = \begin{cases} 1 \text{ if } X_{ij} = k \\ 0 \text{ otherwise.} \end{cases}$$

The item parameters for all test items are

$$\beta = (\vec{\beta}_1, \ldots, \vec{\beta}_j, \ldots, \vec{\beta}_J)^T. \tag{8}$$

The likelihood equation is

$$P(X \mid \theta, \beta) = \prod_{i=1}^{N} P(\vec{X}_i \mid \vec{\theta}_i, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j). \tag{9}$$

Let $P(\theta \mid \lambda)$ be the probability distribution for an examinee population with ability θ, given parameter λ. For example, if θ is assumed to be normally distributed, then $\lambda = (\mu, \sigma)$, where μ is the mean and σ is the variance-covariance matrix. The joint posterior distribution can be written as

$$P(\theta, \beta, \lambda \mid X) \propto P(X \mid \theta, \beta, \lambda) P(\theta \mid \beta, \lambda) P(\beta) P(\lambda) \tag{10}$$

$$= P(X \mid \theta, \beta) P(\theta \mid \lambda) P(\beta) P(\lambda), \tag{11}$$

where

$$P(\theta \mid \lambda) = \prod_{i=1}^{N} P(\vec{\theta}_i \mid \lambda), \tag{12}$$

with the assumption that the examinee's response to an item depends on the examinee's ability and the item parameters.

Markov Chain Monte Carlo Estimation

Multidimensional IRT parameter estimation can be difficult due to numerous additional parameters, when compared to the unidimensional case. An MCMC method samples from the posterior distribution, which avoids taking derivatives in multidimensional space. The parameters (θ, β, λ) are estimated using the Metropolis-Hasting algorithm that samples from the joint posterior $P(\theta, \beta, \lambda \mid X)$. Metropolis-Hastings is a general term for Markov chain simulation methods that are useful for drawing samples from appropriate distributions and then correcting those draws to better target the posterior distribution (Patz & Junker, 1999a, 1999b). A Bayesian formulation of multivariate item response theory was implemented in the computer program BMIRT (Yao, 2003, 2004a, 2004b; Yao & Boughton, 2005b; Yao & Schwarz, in press) that uses the MCMC method to estimate the M-3PL model and M-2PPC model with the ability to conduct both the exploratory and confirmatory approach. Note that all the parameters in this study were estimated using BMIRT.

Simulation Study

Item Parameters

This study simulated examinee responses based on actual dichotomous and polytomous item parameters obtained from a large-scale Grade 8 mathematics assessment, with four objectives or subscales, currently using an OPI subscoring approach. The calibration sample consisted of 10,000 examinees across 60 items for the following objectives:

- Objective or Dimension 1: 15 items, Number Sense and Computational Techniques
- Objective or Dimension 2: 15 items, Algebra, Patterns and Functions
- Objective or Dimension 3: 12 items, Statistics and Probability
- Objective or Dimension 4: 18 items, Geometry and Measurement

The item parameters used in this simulation study were obtained from a four-dimensional confirmatory calibration of the sample described above using BMIRT. A simple structure approach was used, given that the items loaded on only one of the four objectives or subscales based on the content blueprint map assigned in the test development phase. One of the strengths of the MIRT augmentation approach is that it is also well suited for tests with items that load on more than one objective at a time, but the purpose of this research was to compare an MIRT approach to the OPI, which requires each item to load on only one of the objectives at a time. The item parameters and the means of the parameters are listed in Table 1. The first column indicates the item number, the second column indicates item response categories or levels, columns 3 through 6 present the four discrimination parameters, and the last few columns present the difficulty and guessing parameters for M-3PL items or alphas for M-2PPC items. The indeterminacies of the model were solved by fixing the four-dimensional ability $\vec{\theta}$ distribution of the multinormal with mean $(0, 0, 0, 0)$ and the variance-covariance matrix as

$$\sigma = \begin{pmatrix} a_1 & r_1 & r_2 & \cdots \\ r_1 & a_2 & r_3 & \cdots \\ \vdots & \vdots & \ddots & \\ r_4 & r_5 & \cdots & a_4 \end{pmatrix}_{4\times 4} \tag{13}$$

where $r_1 = r_2 = \cdots = 0, a_1 = a_2 = a_3 = a_4 = 1$.

Simulation Conditions

There were three steps involved in the generation and recovery of item and ability parameters:

1. The .0 correlations and a standard variance-covariance matrix were used only to set the scale to obtain the item parameter estimates from the real data. Thus, the correlations in the real data were pushed into the item parameters themselves.

2. Simulated response data were produced using abilities generated from various ability distributions with the fixed item parameters from the real data.

3. To recover the item parameters for these simulated ability distributions using BMIRT, the mean and variance-covariance matrix were set to their true values to fix the indeterminacy of the scale and to keep the estimated item parameters on the same scale as their true parameters, therefore avoiding the need to equate the numerous conditions and replications studied in this research.

Given the above steps, abilities for 1,000, 3,000, and 6,000 examinees were generated from multinormal population distributions with mean $(0, 0, 0, 0)$ and the variance-covariance matrix as in equation (13), with $r = .0, .1, .3, .5, .7$, and .9. The correlations between dimensions were fixed to be equal, in order to reduce the number of conditions to be studied. Each ability, in conjunction with the item parameters in Table 1, was used to simulate examinee responses for both the multiple-choice (M-3PL) and constructed-response (M-2PPC) item types. Twenty different seeds were used to obtain 20 replications across all conditions.

Item Parameter Estimation Procedure

For each varying condition, a four-dimensional confirmatory calibration was conducted using BMIRT, with the number of iterations chosen after reviewing the sampling histories for convergence.

Table 1

True Item Parameters

M-3PL M-2PPC	Score Level	Discrimination ($\vec{\beta}_{2j}$)									
		β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	β_{1j}	β_{3j}	$\beta_{\delta 2j}$	$\beta_{\delta 3j}$	$\beta_{\delta 4j}$	$\beta_{\delta 5j}$
Mean–3PL		1.88	1.80	1.25	1.86	0.91	0.18				
Mean–2PPC		1.21	1.25	1.18	1.05			0.32	0.32	1.05	0.65
1	1	2.53				5.18	0.06				
2	1	2.45				0.55	0.17				
3	3		0.95					0.69	–0.57		
4	1				2.52	3.71	0.11				
5	1			0.80		–2.11	0.19				
6	1			1.26		0.01	0.19				
7	1			1.58		0.54	0.17				
8	1		0.86			–1.09	0.24				
9	4					1.12		–1.34	0.34	–0.56	
10	1		2.28			1.46	0.23				
11	1		1.34			0.58	0.21				
12	3	1.23						2.07	0.46		
13	1				1.93	0.92	0.09				
14	4	0.67						–0.73	2.10	1.60	
15	1	1.70				2.85	0.14				
16	1				1.79	1.41	0.14				
17	1				1.25	0.33	0.19				
18	1	1.58				–0.68	0.14				
19	1				1.27	0.40	0.20				
20	5	1.37						1.48	1.01	2.06	1.12
21	1			1.00		0.17	0.11				
22	1		2.04			2.95	0.23				
23	1		2.35			0.96	0.21				
24	1		0.73			–0.92	0.20				
25	3		1.54					3.43	1.95		
26	1			1.54		–0.86	0.18				
27	1			2.22		3.01	0.09				
28	4			1.15				–1.37	–0.44	0.11	
29	1				1.58	3.55	0.19				
30	3				0.79			2.81	–1.58		
31	1				2.45	1.25	0.23				
32	1	1.44				2.15	0.17				
33	1				3.05	3.25	0.24				
34	4				1.31			–0.29	2.03	2.16	
35	1	2.16				1.28	0.12				
36	1				2.72	4.72	0.21				
37	1	1.85				0.64	0.23				
38	1	2.24				0.09	0.16				
39	1		2.24			1.76	0.20				
40	5				1.14			–2.04	0.15	0.65	0.82
41	1		1.21			1.71	0.14				
42	1				1.66	0.93	0.23				

(continued)

Table 1 (continued)

M-3PL M-2PPC	Score Level	Discrimination ($\vec{\beta}_{2j}$)									
		β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	β_{1_j}	β_{3_j}	$\beta_{\delta 2_j}$	$\beta_{\delta 3_j}$	$\beta_{\delta 4_j}$	$\beta_{\delta 5_j}$
43	3			1.25				1.17	2.24		
44	1				1.50	–0.38	0.20				
45	1	1.11				0.43	0.15				
46	1			1.20		–1.78	0.19				
47	1	2.11				–2.02	0.15				
48	4				0.94			–0.36	–0.92	0.10	
49	1			0.97		–0.18	0.20				
50	3	1.56						2.25	–1.35		
51	1				1.90	1.80	0.06				
52	1		1.92			–1.25	0.18				
53	4		1.24					0.07	0.60	1.15	
54	1	1.53				–0.20	0.22				
55	1				1.79	0.40	0.21				
56	1			0.66		0.13	0.19				
57	1		2.25			0.69	0.16				
58	1				0.54	–1.30	0.20				
59	1		2.52			3.8	0.17				
60	5		1.26					–3.10	–1.17	2.20	–0.007

Note. M-3PL = multidimensional three-parameter logistic model; M-2PPC = multidimensional version of the 2-parameter partial credit model.

Given that it was not feasible to check each individual outcome for convergence in this simulation study, a conservative burn-in length of 10,000 and a total of 30,000 iterations were set with arbitrary starting values for the parameters. A large number of iterations were used to help ensure convergence and thus produce more accurate item parameter estimates across the varying conditions. The approximate running time of BMIRT on a Windows-based Xeon 3.06-GHz desktop machine was 1 hour for a single replication with a sample size of 1,000, 3 hours for sample size of 3,000, and 6 hours for sample size of 6,000.

For the multiple-choice items (M-3PL), the priors were as follows:

$$\beta_{1j} \sim N(\mu_{\beta_{1j}}, \sigma^2_{\beta_{1j}}), \tag{14}$$

$$\log(\beta_{2jl}) \sim N(\log(\mu_{\beta_{2j}}), \sigma^2_{\beta_{2j}}) \tag{15}$$

for $l = 1, \ldots, 4$.

$$\beta_{3j} \sim beta(a, b) \tag{16}$$

and $\mu_{\beta_{1j}} = 0$, $\mu_{\beta_{2j}} = 2$, $a = 6$, $b = 16$, $\sigma_{\beta_{1j}} = 2.5$, and $\sigma_{\beta_{2j}} = 2.5$.

For the constructed-response items (M-2PPC), the priors were taken to be lognormal for each component of β_{2j} and normal for $\beta_{\delta_k j}$. The means and standard deviations of the prior distributions were $\mu_{\beta_{2j}} = 2$, $\mu_{\beta_{\delta_k j}} = 0$, $\sigma_{\beta_{\delta_k j}} = 2.5$, and $\sigma_{\beta_{2j}} = 2.5$, where $k = 2, \ldots, K_j$.

Like unidimensional IRT models, MIRT models have scale indeterminacy, but the probability of a correct response will not change after a linear transformation. Therefore, it was necessary to impose the metric or scale in the estimation. The approach used here to place the estimated parameters onto the same scale as the true parameters, without the need to equate, was to fix the population parameters to multinormal with mean $(0, 0, 0, 0)$ and the variance-covariance matrix, as in Equation (13), with diagonal element $a = 1$ and correlations $r = 0, .1, .3, .5, .7, .9$, accordingly. Even though in practice, the true correlations are unknown and the correlations may not be the same, they can be approximated by finding the correlations between the subtotal scores for each objective or the correlations between the estimated unidimensional objective-level scores. Yao (2005) has found that the item parameter recoveries from both of the above options produced similar results compared to using the true correlations in conditions in which the correlations were varied across dimensions. Specifically, it was found that as long as the fixed correlations were smaller than the true correlations, the resulting estimated item parameters were close to their true item parameters. Another option to convert the estimated item parameters onto the same scale as their true item parameters would be to use a multidimensional scale-linking approach based on a generalization of the Stocking and Lord (1983) procedure, as demonstrated in Yao (2005) and Yao and Boughton (2005a).

Subscale Score Estimation Procedures

Percentage of Number Correct (NC)

The percentage of the number correct was the simplest model for obtaining a subscale score. It was computed for each objective as the number of points obtained by a student on the set of items in that objective divided by the total number of points obtainable on the items for that objective, expressed as a percentage.

Unidimensional Objective-Level Bayesian Score (UIRTOJSS)

Unidimensional parameter estimation from BMIRT for each objective, based only on the items in that objective, was conducted for conditions $r = 0, .1, .5, .9$ for a sample size of 6,000. The purpose of comparing a unidimensional with a multidimensional estimation approach was to substantiate how much error was introduced when using a multidimensional Bayesian modeling approach.

Multidimensional Pattern Scale Score (MIRTPSS)

To compare MCMC to MML estimation, scale scores were obtained using a maximum likelihood (ML) grid search approach for finding the value of theta that maximizes the likelihood function. Item parameter estimates were put onto a different scale score metric in order to use the software, which entails two steps. The first step was to find reasonable multipliers and additives to apply to a standard normal population distribution metric (0/1 metric) for the item parameter estimates. The second step was to establish a range of allowable scale scores—that is, set the highest and lowest obtainable scale scores (referred to as the HOSS and LOSS, respectively).

For each examinee i, with responses X_i, the likelihood function that needs to be maximized is

$$P(\vec{X}_i \mid \vec{\theta}_i, \beta) = \prod_{j=1}^{J} P_{ij}(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j). \tag{17}$$

In the case of simple structure, each item contributes to one and only one dimension (or objective) and can be decomposed to

$$P(\vec{X}_i \mid \vec{\theta}_i, \beta) = \prod_{l=1}^{D} \prod_{j \in O_l} P_{ij}(X_{ij} \mid \theta_{il}, \vec{\beta}_j), \tag{18}$$

where O_l contains the items that contribute to the objective l and $l = 1, \ldots, D$. Therefore, maximizing

$$P(\vec{X}_i \mid \vec{\theta}_i, \beta), \tag{19}$$

to obtain estimation for

$$\vec{\theta}_i = (\theta_{i1}, \ldots, \theta_{iD}), \tag{20}$$

is equivalent to maximizing

$$\prod_{j \in O_l} P_{ij}(X_{ij} \mid \theta_{il}, \vec{\beta}_j), \tag{21}$$

to obtain θ_{il} for $l = 1, \ldots, D$. Given the current unavailability of ML estimation software for the M-3PL and M-2PPC model combination implemented in this study, the ML estimation procedure described above was used here to obtain the four ability estimates (i.e., four dimensions) separately by separating the four-dimensional item parameter estimates from BMIRT by objective, using the discrimination parameters from each dimension as that dimension's discrimination. There was only one difficulty parameter across the four dimensions and this single difficulty parameter was used for each of the four dimensions, with that dimension's discrimination parameter resulting in a unidimensional parameterization. The item parameters were transformed from the β_j (0/1 metric) to the β_j^* scale score metric, $j = 1, \ldots, J$, through the multiplicative and linear constants $M1 = 50$, $M2 = 500$ as follows.

Let

$$\vec{\theta}_i^* = \vec{\theta}_i A^T + \vec{B}, \tag{22}$$

with matrix A as in equation (13), with $r_1 = r_2 = \cdots = 0$, $a_1 = \cdots = M1$, and

$$\vec{B} = (M2, \ldots, M2). \tag{23}$$

For M-2PPC items,

$$\vec{\beta}_{2j}^* = \vec{\beta}_{2j} A^{-1}, \tag{24}$$

and

$$\beta_{\delta_k j}^* = \beta_{\delta_k j} + \beta_{2j} A^{-1} \vec{B}^T, \tag{25}$$

then

$$P_{ijk}(\vec{\theta}_i^*, \vec{\beta}_j^*) = P_{ijk}(\vec{\theta}_i, \vec{\beta}_j) \tag{26}$$

For M-3PL items, the transformation is

$$\vec{\beta}_{2j}^* = \frac{\vec{\beta}_{2j} A^{-1}}{1.7}, \tag{27}$$

$$\beta_{1j}^{*} = M2 + M1\beta_{1j}MID^{-1}, \tag{28}$$

$$MID = \vec{\beta}_{2j} \odot \vec{\beta}_{2j}^{T}, \tag{29}$$

$$\beta_{3j}^{*} = \beta_{3j}, \tag{30}$$

then

$$-\vec{\beta}_{2j}^{*} \odot \vec{\theta}_{i}^{*T} + \beta_{1j}^{*} = -\vec{\beta}_{2j} \odot \vec{\theta}_{i}^{T} + \beta_{1j}. \tag{31}$$

The highest and lowest obtainable scale scores, HOSS and LOSS, were set by $\text{HOSS} = 4M1 + M2$ and $\text{LOSS} = -4M1 + M2$. The scale score estimates for each student on each objective were obtained through the MIRTPSS maximum likelihood grid search algorithm, as previously described. Note that the model and the metric of the parameters obtained from BMIRT were specified in equations (2) and (4), and the transformation above was necessary for the software to obtain MIRTPSS.

Bayesian Multidimensional Scale Score (BMIRTSS)

Ability $\vec{\theta}_i$, $i = 1, \ldots, N$ and population distribution parameters were obtained from running BMIRT by fixing the item parameters estimated from the previous section. This Bayesian score estimation was in standard multinormal metric. Scale scores $\vec{\theta}_i^{*}$ in the same metric as BMIRTPSS were obtained by

$$\vec{\theta}_i^{*} = \vec{\theta}_i A^{T} + \vec{B}, \tag{32}$$

where A and $\vec{B}$ were defined as the above section.

Bayesian Multidimensional Domain Score (BMIRTDS)

Domain scores from the four-dimensional BMIRT solution were obtained as follows: For student i, the four domain scores $(D_1, \ldots, D_4)$, were obtained by

$$\frac{D_l = \sum_{j \in O_l, j \in M-2PPC} \sum_{k=1}^{K_j} (k-1) P_{ijk}(x_{ij} = k-1 \mid \vec{\theta}_i, \vec{\beta}_j) + \sum_{j \in O_l, j \in M-3PL} P_{ij1}(x_{ij} = 1 \mid \vec{\theta}_i, \vec{\beta}_j)}{\sum_{j \in O_l, j \in M-2PPC} (K_j - 1) + \sum_{j \in O_l, j \in M-3PL} 1}, \tag{33}$$

where O_l contains the items that contribute to the objective l and $l = 1, 2, 3, 4$. The denominator of equation (33) was the total score points in the test, whereas the numerator was the summation of the expected points for each item. The polytomous items were specified as M-2PPC and multiple choice as M-3PL. D_l was the expected percentage of points for the test for objective l. The true domain scores were obtained by using the true parameters for $\vec{\theta}_i$ and $\vec{\beta}_j$, whereas BMIRTDS estimates were obtained by using $\vec{\theta}_i$ and $\vec{\beta}_j$ from the BMIRT four-dimensional calibration. Note that BMIRTDS is in the same metric as the NC scores.

Objective Performance Index (OPI)

To obtain OPI estimates, BMIRT unidimensional item parameter calibrations were run with 30,000 iterations, with the first 10,000 as burn-in. D_l was the mean of the prior distribution for

objective l, and it was defined the same as in equation (33), except that the parameters were unidimensional IRT estimates. $OPI = (OPI_1, \ldots, OPI_4)$ were obtained as follows: OPI_l was defined as the mean of the posterior distribution (Yen, 1987) $P(D_l \mid Y_l)$, which is

$$P(D_l \mid Y_l) \propto P(Y_l \mid D_l) P(D_l), \tag{34}$$

where $l = 1, \ldots, 4$, and

$$Y_l = \sum_{j \in O_l} X_j \tag{35}$$

was the observed score for objective l, where X_j was the response of an examinee (i, which represents examinee i in X_{ij}, was dropped for convenience) to an item j. Assume

$$Y_l \mid D_l \sim binomial(J, D_l) \tag{36}$$

and

$$D_l \sim \beta(r_l, s_l), \tag{37}$$

then

$$D_l \mid Y_l \sim \beta(p, q), \tag{38}$$

where $p_l = r_l + x$, $q_l = s_l + J - x$, and x was the observed score of Y_l. r and s can be obtained from the following two equations:

$$\mu(\hat{D}_l \mid \theta) = \frac{r_l}{r_l + s_l} \tag{39}$$

and

$$\sigma^2(\hat{D}_l \mid \theta) = \frac{r_l s_l}{(r_l + s_l)^2 (r_l + s_l + 1)}. \tag{40}$$

Therefore,

$$OPI_l = \frac{p_l}{p_l + q_l}. \tag{41}$$

For more detailed information about OPI, see Yen (1987). Given that the OPI borrows information from the total test score to help stabilize the objective subscores, it has been found to be more reliable than a simple number-correct subscore (Yen, 1987).

Recovery Criteria

A root mean square error (RMSE) was computed for each parameter and then averaged across parameters. Let f_{true} be the true parameter. Let f_j be the estimated parameter from sample j. RMSE was calculated by

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (f_j - f_{true})^2}, \tag{42}$$

where n is the number of replications.

Table 2

Root Mean Square Error (RMSE) for Multiple-Choice Item Parameter Recovery

RMSE for Sample Size of 1,000						
	Parameters					
Correlation	β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	β_{1j}	β_{3j}
.0	0.442	0.353	0.322	0.350	0.412	0.049
.1	0.443	0.349	0.304	0.346	0.412	0.049
.3	0.455	0.342	0.290	0.360	0.424	0.049
.5	0.434	0.336	0.240	0.350	0.430	0.048
.7	0.416	0.304	0.252	0.348	0.427	0.048
.9	0.385	0.299	0.271	0.347	0.420	0.048
RMSE for Sample Size of 3,000						
	Parameters					
Correlation	β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	β_{1j}	β_{3j}
.0	0.243	0.212	0.150	0.226	0.250	0.036
.1	0.245	0.208	0.142	0.223	0.248	0.036
.3	0.250	0.197	0.142	0.213	0.248	0.036
.5	0.246	0.186	0.137	0.214	0.251	0.037
.7	0.230	0.176	0.147	0.212	0.251	0.036
.9	0.215	0.173	0.154	0.199	0.247	0.036
RMSE for Sample Size of 6,000						
	Parameters					
Correlation	β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	β_{1j}	β_{3j}
.0	0.168	0.158	0.112	0.163	0.191	0.030
.1	0.167	0.160	0.108	0.163	0.192	0.030
.3	0.166	0.158	0.115	0.159	0.191	0.029
.5	0.159	0.154	0.105	0.155	0.190	0.029
.7	0.149	0.132	0.108	0.165	0.182	0.028
.9	0.139	0.132	0.109	0.147	0.169	0.027

Results

The unidimensional and multidimensional item parameters for the 3PL and 2PPC models were estimated by MCMC, with ability estimation by MCMC and ML. As previously mentioned, an ML procedure for ability estimation using the MCMC item parameter estimates was studied, as it is important to establish the accuracy of the MCMC approach, when compared to a solution strategy that has already been well established.

Item Parameter Recovery

The RMSEs for the MC items can be found in Table 2 and the CR RMSEs in Table 3, by sample size and correlation group. Note that although the tables show RMSEs for all parameters, they are

Table 3

Root Mean Square Error (RMSE) for Constructed-Response Item Parameter Recovery

RMSE for Sample Size of 1,000								
	Parameters							
Correlation	β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	$\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$	$\beta_{\delta_4 j}$	$\beta_{\delta_5 j}$
.0	0.114	0.101	0.108	0.098	0.148	0.150	0.161	0.177
.1	0.115	0.107	0.112	0.095	0.150	0.163	0.165	0.188
.3	0.114	0.112	0.099	0.086	0.154	0.170	0.160	0.192
.5	0.112	0.114	0.091	0.085	0.153	0.167	0.168	0.195
.7	0.108	0.110	0.092	0.085	0.144	0.172	0.182	0.211
.9	0.100	0.093	0.088	0.077	0.145	0.183	0.175	0.194
RMSE for Sample Size of 3,000								
	Parameters							
Correlation	β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	$\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$	$\beta_{\delta_4 j}$	$\beta_{\delta_5 j}$
.0	0.070	0.069	0.051	0.054	0.081	0.087	0.089	0.105
.1	0.072	0.069	0.059	0.054	0.079	0.089	0.088	0.110
.3	0.074	0.070	0.059	0.049	0.079	0.092	0.088	0.123
.5	0.072	0.068	0.055	0.047	0.078	0.092	0.088	0.106
.7	0.071	0.067	0.054	0.041	0.079	0.095	0.094	0.107
.9	0.065	0.057	0.056	0.039	0.082	0.090	0.084	0.111
RMSE for Sample Size of 6,000								
	Parameters							
Correlation	β_{2j1}	β_{2j2}	β_{2j3}	β_{2j4}	$\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$	$\beta_{\delta_4 j}$	$\beta_{\delta_5 j}$
.0	0.059	0.045	0.036	0.040	0.056	0.064	0.077	0.080
.1	0.058	0.049	0.040	0.038	0.055	0.062	0.077	0.076
.3	0.057	0.047	0.042	0.037	0.055	0.064	0.073	0.079
.5	0.054	0.045	0.040	0.031	0.055	0.059	0.069	0.069
.7	0.050	0.043	0.038	0.029	0.058	0.062	0.074	0.075
.9	0.048	0.037	0.035	0.028	0.058	0.063	0.067	0.070

not all on the same metric and cannot all be directly compared. Comparisons between sample sizes and populations for each parameter can be made. Overall, there was a larger reduction in RMSE when moving from a sample size of 1,000 to 3,000, with a less dramatic reduction in error when increasing the number of examinees to 6,000. As the correlations between dimensions increased, RMSEs decreased, as more information was borrowed from the other dimensions. The RMSE values for the CR items also decreased as the sample size increased, but the improvement in accuracy as the correlation increased was not evident.

Ability Parameter Recovery

The examinee ability parameters (BMIRTSS) and population parameters were both estimated from BMIRT by fixing the item parameters from the initial BMIRT run. A comparison of this

Table 4

Average Root Mean Square Error Proficiency Recovery for BMIRTSS, MIRTPSS, and UIRTOJSS (range: 300-700)

	BMIRTSS											
	Dimension 1			Dimension 2			Dimension 3			Dimension 4		
	Sample Size			Sample Size			Sample Size			Sample Size		
Correlation	1,000	3,000	6,000	1,000	3,000	6,000	1,000	3,000	6,000	1,000	3,000	6,000
.0	20	20	20	21	21	21	23	23	23	20	20	20
.1	20	20	20	21	21	21	23	23	23	19	20	20
.3	20	20	20	21	20	20	23	23	22	19	19	19
.5	20	19	19	20	20	20	22	22	22	19	19	19
.7	18	18	18	19	18	18	20	20	20	18	18	18
.9	16	15	15	16	16	16	17	16	16	15	15	15
	MIRTSS											
.0	44	39	38	29	29	29	33	32	32	29	28	29
.1	44	40	38	29	29	29	33	32	32	30	29	29
.3	44	40	38	29	29	29	33	32	32	30	29	29
.5	43	40	38	30	29	29	33	32	32	29	29	29
.7	43	39	38	30	29	29	33	32	32	30	29	29
.9	43	39	37	30	29	29	33	32	32	30	29	28
	UIRTOJSS											
.0			20			21			24			21
.1	20		20	21		21	23		23	20		20
.5			20			21			23			20
.9			20			21			23			20

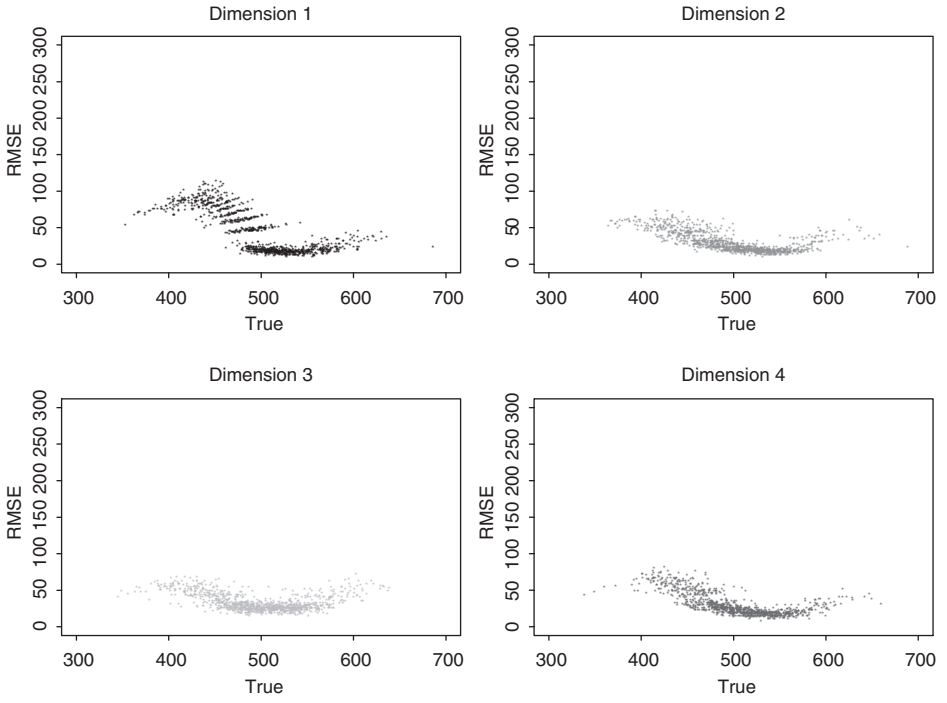
Note. BMIRTSS = multidimensional item response theory Bayesian subscale scores; MIRTPSS = maximum likelihood item response theory pattern subscale scoring approach; UIRTOJSS = unidimensional item response theory objective-level Bayesian scoring approach.

Bayesian ability estimation (BMIRTSS) to an ML approach (MIRTPSS) was implemented along with a Bayesian unidimensional approach (UIRTOJSS). Table 4 displays the RMSE values for BMIRTSS, MIRTPSS, and UIRTOJSS on a scale ranging from 300 to 700 for each of the four subscales. The recovery of BMIRTSS was better than the recovery of MIRTPSS across all conditions. BMIRTSS and UIRTOJSS had similar RMSEs when the correlation was low (e.g., $r = .0$ or .1), but as the correlation increased, the RMSEs for BMIRTSS decreased, whereas for UIRTOJSS, the RMSEs remained the same. The correlation had no impact on the recovery of MIRTPSS and UIRTOJSS and was a result of the fact that they did not use that correlation information in ability estimation.

Figure 1 displays four plots for MIRTPSS (ML approach) recovery for a sample size of 1,000 and population correlation $r = .0$, with the RMSEs on the y-axis and true scale scores on the x-axis. These graphs show that the RMSEs were largest for lower ability examinees, especially in Dimension 1, and may be a direct result of the lower amounts of item information at the lower ability levels.

Figure 2 shows the RMSE of BMIRTSS for correlations $r = .0$, .5, and .9, with a single sample size of 1,000, as the same pattern was observed for both the 3,000-examinee and 6,000-examinee

Figure 1
RMSE for MIRTPSS Estimation for Correlation $r = .0$



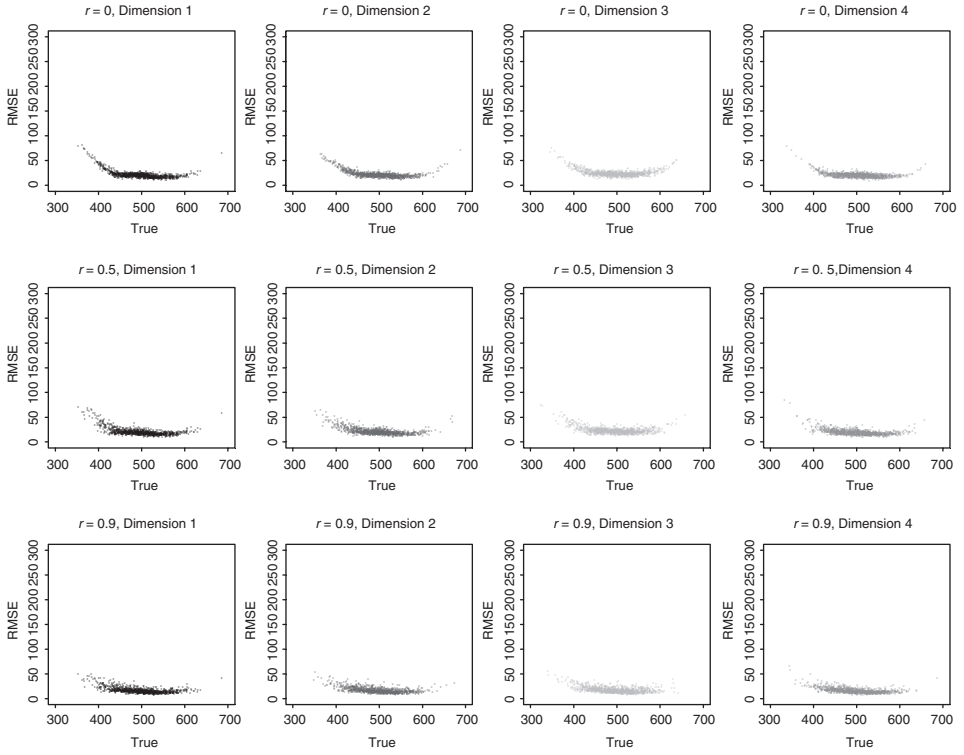
Note. RMSE = root mean square error; MIRTPSS = maximum likelihood item response theory pattern subscale scoring approach.

conditions. In comparison with the pattern scoring by MIRTPSS (Figure 1), the error in ability estimation by BMIRTSS was greatly reduced. As the correlation increased, the RMSEs for BMIRTSS decreased, as seen in Table 4 (averaged over examinees). However, Figure 2 shows that it was in fact the RMSEs for the lower and higher ability examinees that were decreased, with a slight increase for the average-ability examinees.

In estimating ability, the observed correlations between MIRT-based scale score estimates (BMIRTSS) were slightly larger (although not to a great extent) than the true score correlations, even though the estimated population distribution mean and variance-covariance matrices were very close to their true values. The correlated errors might be the cause of this increase. As mentioned, correlations between subscales for the true values in this study were $(.0, .1, .3, .5, .7, .9)$, whereas observed correlations between estimated BMIRTSS were slightly larger at $(.0, .12, .40, .65, .85, .98)$. In addition, Bayesian estimation tended to underestimate the higher ability parameters while overestimating the lower ability parameters. Thus, there tends to be shrinkage in the variance of the estimated examinees' abilities. In fact, the variance components for the true value of 1 were estimated at .69, .69, .71, .73, .77, and .84. Note that this phenomenon of Bayesian estimation exists even for unidimensional IRT models.

Table 5 displays the RMSE values for BMIRTDS and OPI on the percentage correct scale across the four subscales. The recovery of BMIRTDS was better than OPI across all conditions. Overall, the RMSE for OPI became smaller as the correlation increased, producing similar results

Figure 2
RMSE for BMIRTSS Score Estimation for Correlations of .0, .5, and .9, With a Sample Size of 1,000



Note. RMSE = root mean square error; BMIRTSS = multidimensional item response theory Bayesian subscale scores.

to BMIRTDS, at a correlation of .9. As expected, as the correlation increased, the RMSEs for both BMIRTDS and OPI decreased and are a direct result of the utilization of the relationships among the various subscales by BMIRTDS and the total test score information by OPI. Note that an increase in sample size did not improve the recovery rates of ability.

Figure 3 shows a comparison between OPI and the true domain score in percentage correct units, whereas Figure 4 shows a comparison between BMIRTDS and the true domain score for a sample size of 1,000 and correlations $r = .0$, .5, and .9. It is clear that the correlations among the subscores did affect the estimation accuracy of both OPI and BMIRTDS. As the correlation between the subscores increased, OPI and BMIRTDS errors decreased and became more similar to each other. For the .0 correlation condition, OPI and BMIRTDS produced overestimates for lower ability examinees, but OPI had more error than BMIRTDS for correlations ranging from .0 to .5, with similar results at a correlation of .9.

Classification Accuracy

To establish performance levels for classification recovery, cut-scores were first defined by splitting up the true distribution, such that 20% of the students were Below Basic, 20% were Basic,

Table 5

Average Root Mean Square Error Proficiency Recovery for BMIRTDS and OPI (range: 0-100)

BMIRTDS												
	Subscale											
	Dimension 1			Dimension 1			Dimension 1			Dimension 1		
	Sample Size			Sample Size			Sample Size			Sample Size		
Correlation	1,000	3,000	6,000	1,000	3,000	6,000	1,000	3,000	6,000	1,000	3,000	6,000
.0	7	7	7	8	8	8	9	9	9	7	7	7
.1	7	7	7	8	8	8	9	9	9	7	7	7
.3	7	7	7	7	7	7	9	9	9	7	7	7
.5	7	7	7	7	7	7	8	8	8	7	7	7
.7	7	7	7	7	7	7	8	8	8	7	7	7
.9	6	6	6	6	6	6	6	6	6	6	6	6
OPI												
	Subscale											
	Dimension 1			Dimension 1			Dimension 1			Dimension 1		
	Sample Size			Sample Size			Sample Size			Sample Size		
Correlation	1,000	3,000	6,000	1,000	3,000	6,000	1,000	3,000	6,000	1,000	3,000	6,000
.0	10	10	10	10	11	11	12	12	12	9	9	10
.1	10	10	10	10	10	10	12	12	12	9	9	9
.3	10	9	9	10	10	10	11	11	11	9	9	9
.5	9	9	9	9	9	9	11	11	10	9	9	9
.7	8	8	8	8	9	8	10	10	10	8	8	8
.9	7	7	7	7	7	7	8	8	8	7	7	7

Note. BMIRTDS = multidimensional item response theory Bayesian domain subscale scores; OPI = objective performance index.

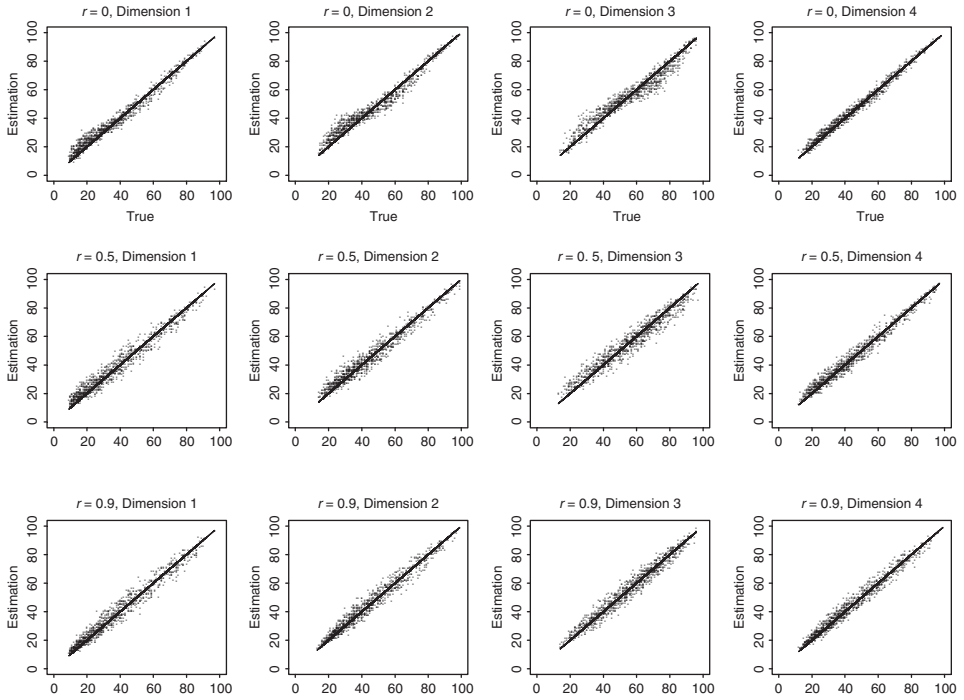
40% were Proficient, and 20% were Advanced. These performance levels are defined by the scale scores 457, 487, and 542, respectively.

BMIRTSS and MIRTPSS were both expressed in a three-digit latent score metric, whereas NC, OPI, and BMIRTDS were all in the number-correct scale. Thus, the number-correct cut-scores were obtained by using equation (33), with the scale score cut-scores as $\vec{\theta}_i$ and with the true item parameters as $\vec{\beta}_j$, yielding the following cut-scores:

- Subscale 1: 19, 27, 57
- Subscale 2: 28, 37, 64
- Subscale 3: 38, 51, 75
- Subscale 4: 26, 36, 63

Classification errors were of two kinds. False-positive (FP) errors occurred when the estimated category was higher than the true category, and false-negative (FN) errors occurred when the

Figure 3
Comparison of OPI With True Domain Score for Correlations of .0, .5, and .9, With a Sample Size of 1,000



Note. OPI = objective performance index.

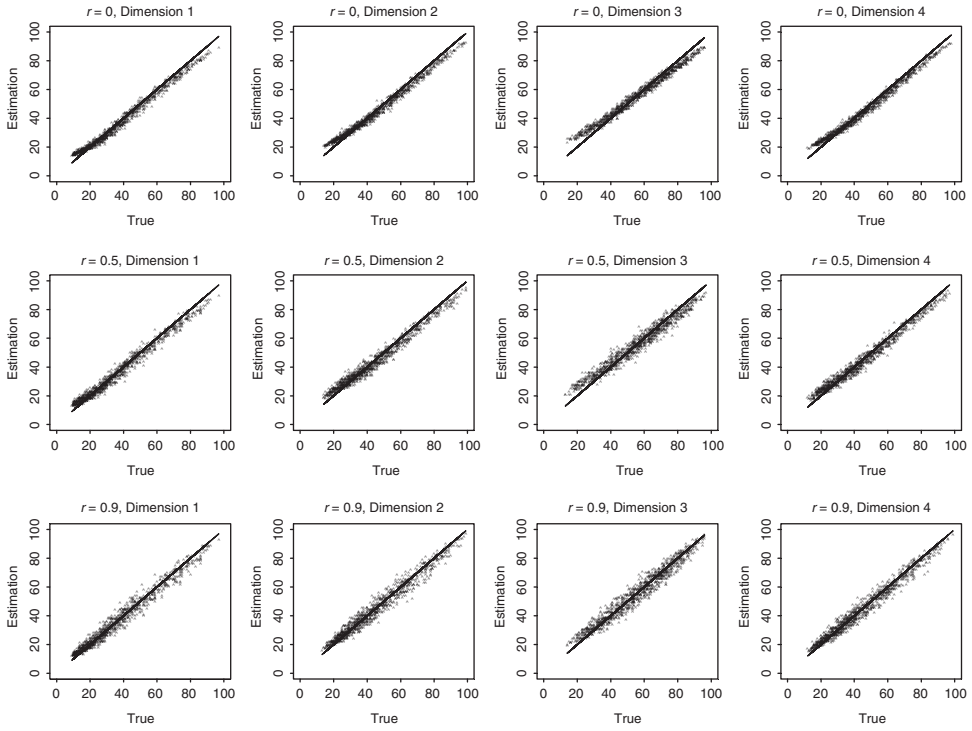
estimated category was lower than the true category. Figure 5 shows these misclassification rates for a sample size of 6,000, averaged across subscales for each of the correlation groups.

NC subscores had the largest errors, regardless of correlation group and subscale, and were not displayed in this graph, given the larger error rates (approximately 65% misclassification across the three cut-scores). MIRTPSS had the lowest misclassification rates on average, but the rates varied greatly. MIRTPSS consistently had the lowest FP rate but some of the highest FN rates for all subscales and correlation groups. BMIRTSS and BMIRTDS error rates were found to be very similar to each other, as they should be. As the correlations increased across the dimensions, the average error rates for BMIRTSS and BMIRTDS became close to the OPI rates. Predictably, as the correlation among the dimensions decreased, the error rates for OPI increased, and BMIRTSS and BMIRTDS classification errors decreased. For the .9 correlation group, OPI, BMIRTDS, and BMIRTSS produced very similar results, but BMIRTDS and BMIRTSS clearly outperformed OPI as the correlation moved toward .0.

Discussion and Conclusion

The demand for more formative assessments to be used for in-class diagnostic purposes implies a need for a more fine-grained analysis at the subscale level. This research presents an MCMC

Figure 4
Comparison of BMIRTDS With True Domain Score for Correlations of .0, .5, and .9, With a Sample Size of 1,000



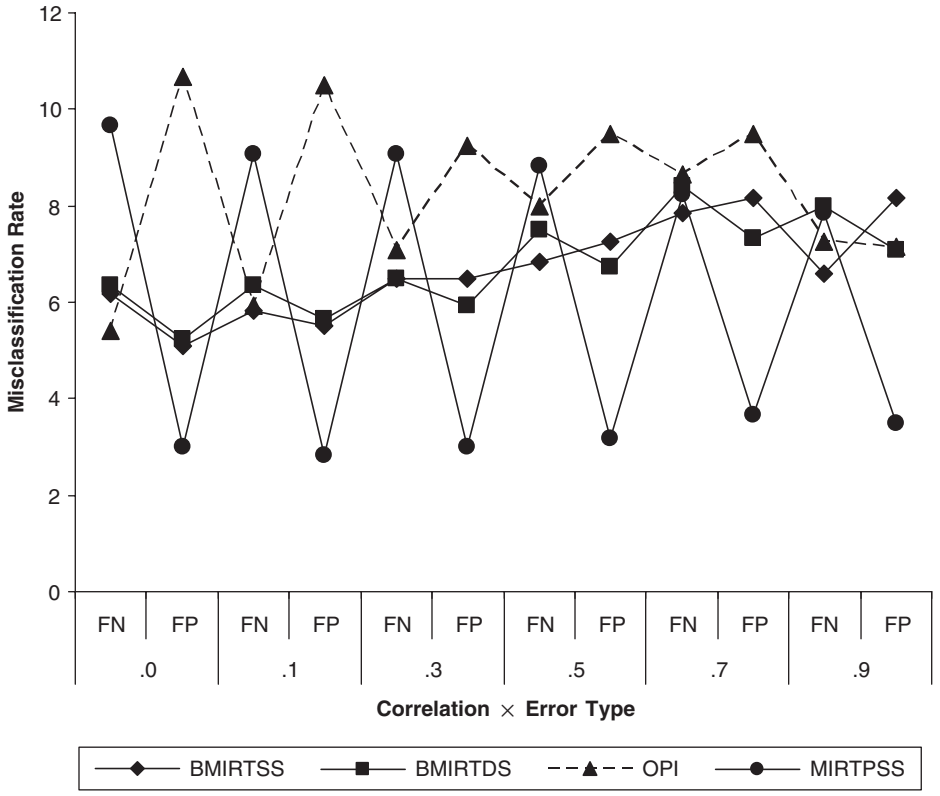
Note. BMIRTDS = multidimensional item response theory Bayesian domain subscale scores.

multidimensional dichotomous and polytomous IRT modeling approach that can be used for subscale score reporting as well as subscale proficiency estimation and classification that leads toward a more diagnostic approach. The dimensional structures were associated with the content objective information in order to report objective scale scores and performance levels. Overall, parameter estimates (i.e., item, ability, objective-level ability, and classification estimation) from BMIRT were well recovered.

Within any simulation study, it is always difficult to quantify what an acceptable level of accuracy is. One way is to find the lower bound, where the recovery has higher error, and then increase the sample size until the change across sample sizes does not produce any more accuracy. Given these guidelines and the poor recovery rates found prior to this study for a sample size smaller than 1,000, the 1,000 sample size conditions were chosen as the lower bound; however, great improvement in item parameter estimation was made when moving to 3,000, with fewer gains seen when moving to 6,000 examinees. Thus, a sample size of at least 3,000 seems to be appropriate for the types of conditions studied here. Note that ability estimation was not greatly improved with an increase in sample size.

The amount of error in item parameter estimation using BMIRT decreased as the correlation increased, with greater increases found when the sample size was increased. In this study, BMIRTSS performed better than MIRTPSS pattern scoring because information from the other

Figure 5
Misclassification rates of BMIRTSS, BMIRTDS, OPI, and MIRTPSS
Across Error Type and Correlation Group for Sample Size of 6,000



Note. FN = false-negative errors; FP = false-positive errors; BMIRTSS = multidimensional item response theory Bayesian subscale scores; BMIRTDS = multidimensional item response theory Bayesian domain subscale scores; OPI = objective performance index; MIRTPSS = maximum likelihood item response theory pattern subscale scoring approach.

subscales (i.e., dimensions) was used, whereas MIRTPSS estimated the subscores using only the item parameters in that subscale. UIRTOJSS (i.e., unidimensional objective-level scoring) estimation and BMIRTSS performed similarly when the correlation was low or close to .0, but as the correlation increased, BMIRTSS performed better, whereas UIRTOJSS remained the same. The recovery of the population parameters using the OPI from the unidimensional solution performed better as the correlation across the subscales increased. It was found that a correlation of .9 (.8 may also work but was not a condition here) is needed for unbiased subscale estimation when using OPI, which borrows information from the total test score. Overall, OPI performed reasonably well when the correlations between populations were high; however, BMIRTDS and BMIRTSS did a much better job of recovering the true objective subscores across all correlation conditions.

For the classification accuracy conditions, OPI performed well or as well as BMIRTDS and BMIRTSS when the correlations were high. As the correlation decreased, the classification errors for BMIRTDS and BMIRTSS decreased, but it increased for OPI. On average, MIRTPSS performed well, whereas the FN rates were as high as the average of the OPI rates, but the OPI FP rates were

similar to the BMIRT classification results. BMIRTDS and BMIRTSS's FN and FP rates were very similar to each other and were consistent across the correlation groups, albeit slightly worse for the higher correlation groups. The reason for this apparent discrepancy was that overall, the RMSEs did decrease; however, as seen in Figure 2, the lower and upper ability examinees' RMSEs were the ones that decreased, whereas the average-ability examinees' errors actually increased. Thus, the classification errors increased as middle-ability examinee scores became biased, most likely due to the correlated errors.

The examinees' abilities (BMIRTSS) were estimated by MCMC through fixing the item parameters in this study. The drawback of this approach is that two examinees with the same response pattern could yield different ability estimates. One potential solution would be to create a Bayesian multidimensional ability estimation software program based on pattern scoring using the item parameter estimates produced from BMIRT. In addition, the MCMC procedure in this study used the same priors for all the examinees, resulting in higher correlations between the estimated abilities (i.e., bias) and shrinkage of variance. A potential remedy would be to use person-specific priors or less informative priors.

Dichotomous multidimensional models have been employed in many different settings; however, polytomous MIRT models for tests with mixed formats are relatively new. The MCMC solution presented here is a first attempt at increasing the estimation accuracy for subscales, improving on the OPI approach when subscale scores are not highly correlated. Another important strength of this approach is that multidimensional equating can be done at the subscale level (Yao, 2004a, 2004b; Yao & Boughton, 2005a), thus making subscores across forms and populations comparable. MCMC methodology is still in its infancy within the area of psychometrics; however, with the growing body of research, its future looks promising. It should be noted that priors can have an impact on parameter estimation; thus, real data applications should use uniform priors and only use more specific priors when that information is available.

More research in this area of diagnostic subscale proficiency estimation using dichotomous and polytomous MIRT models is warranted. Although the simulation study in this article used simple structure, MIRT can also be applied to subscales with complex structure (e.g., Boughton, Yao, & Lewis, 2006). Future research should be conducted to assess how many items within a given subscale are needed to produce accurate and stable subscale ability estimates for both simple and complex structured approaches, while borrowing information from the other subscales.

References

Ackerman, T. A. (1994a). Creating a test information profile in a two-dimensional latent space. *Applied Psychological Measurement*, *18*, 257-275.

Ackerman, T. A. (1994b). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *20*, 309-310.

Ackerman, T. A., & Davey, T. C. (1991, April). *Concurrent adaptive measurement of multiple abilities.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-561.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*, 197-211.

Boughton, K. A., Yao, L., & Lewis, D. (2006, April). Reporting diagnostic subscale scores for tests composed of complex structure. In K. A. Boughton & L. Yao (Organizers), *Improving diagnostic subscore estimation and classification.* Symposium presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Davey, T., & Hirsh, T. M. (1991, April). *Examinee discrimination as measurement properties of multidimensional tests.* Paper presented at the

annual meeting of the American Educational Research Association, Chicago.

Fraser, C. H. (1987). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models for latent trait theory [Computer program]. Armidale, New South Wales, Australia: Center for Behavioral Studies, the University of New England.

Kahraman, N., & Kamata, A. (2004). Increasing the precisions of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, *28*, 407-426.

Lane, S. (2005, April). *Status and future directions for performance assessments in education*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Muraki, E. (1999). POLYFACT Version 2 [Computer program]. Princeton, NJ: Educational Testing Service.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-346.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401-412.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores: ''Borrowing Strength'' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*, 255-275.

Waller, N. (2002). *MicroFACT: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems*. St. Paul, MN: Assessment Systems Corporation.

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

Wilson, D., Wood, R., & Gibbons, R. D. (1987). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer program]. Mooresville, IN: Scientific Software.

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2004a). *Bayesian multivariate item response theory and BMIRT software*. 2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]. Alexandria, VA: American Statistical Association.

Yao, L. (2004b). LinkMIRT: Linking of multivariate item response models [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2005). *An investigation of scaling options in estimating parameters for multidimensional item response theory model*. Manuscript submitted for publication.

Yao, L., & Boughton, K. A. (2005a). *Multidimensional linking for test with mixed item type*. Manuscript submitted for publication.

Yao, L., & Boughton, K. A. (2005b). *Multidimensional parameter recovery from BMIRT and NOHARM*. Manuscript submitted for publication.

Yao, L., & Schwarz R. (in press). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement*.

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

Acknowledgments

The authors express their gratitude to the editor and anonymous reviewers for their comments and suggestions.

Author's Address

BMIRT and LinkMIRT are available for research purposes only. Send requests for software or reprints or further information to Lihua Yao, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940; e-mail: Lihua_Yao@ctb.com.