Running head: OVERALL SCORES AND DOMAIN SCORES

Reporting Valid and Reliable Overall Scores and Domain Scores

Lihua Yao

Defense Manpower Data Center

DoD Center Monterey Bay

Lihua.Yao@osd.pentagon.mil

Abstract

In educational assessment, overall scores obtained by simply averaging a number of domain scores are sometimes reported. However, simply averaging the domain scores ignores the fact that different domains have different score points, that scores from those domains are related, and that at different score points, the relationship between overall score and domain score may be different. In order to report reliable and valid overall scores and domain scores, I investigated the performance of four methods using both real and simulation data: (a) the unidimensional IRT model; (b) the higher-order IRT model, which simultaneously estimates the overall ability and domain abilities; (c) the multidimensional IRT (MIRT) model, which estimates domain abilities and uses the maximum information method to obtain the overall ability; and (d) the bifactor general model. My findings suggest that the MIRT model not only provides reliable domain scores, but also produces reliable overall scores. The overall score from the MIRT maximum information method has the smallest standard error of measurement. In addition, unlike the other models, there is no linear relationship assumed between overall score and domain scores. Recommendations for sizes of correlations between domains and the number of items needed for reporting purposes are provided.

Key words: BMIRT, MCMC, Multidimensional Item Response Theory, Multidimensional Information Function, Overall Score, Domain Score.

Reporting Valid and Reliable Overall Scores and Domain Scores

One example of an assessment in which both domain scores and a composite score is reported is for the $TerraNova^{TM}$, a product of CTB/McGraw-Hill, which measures five main content areas: Mathematics, Reading, Language, Science, and Social Studies. Scale scores are reported for each of the five content areas and, in addition, the average of these scale scores is reported as a composite score. When both content area scores and a composite are reported it is very important that both types of scores are reliable. The issue of what type of composite scores to report has not been addressed through research, and in practice, the simple averaging method has almost always been adopted (CTB/McGraw-Hill, 2001, 2008). However, simply averaging the scores from different content areas ignores the fact that (a) different content areas have different maximum raw score points; (b) scores from those content areas are related; and (c) at different score points, the relationship between composite scores and content scores may be different.

On the other hand, within a single content or subject, many state assessments require the testing company to report not only subject/content scores, but also objective-level scores or domain scores for that subject/content, so that teachers can consider strengths and weaknesses of students in the content areas. For example, in a 60-item mathematics test, 4 domains–algebra, computation, geometry, and probability–are covered and the scores for each domain are reported. In practice, this can be achieved by using unidimensional IRT model to obtain the item parameter estimates. Overall ability can be estimated from these unidimensional item parameter estimates; each domain score can be obtained from the items contributing to that domain (if there are enough items in that objective). Correlation information between domains is typically ignored when this is done.

The two situations described above are similar and both boil down to two questions: (a) what is the relation between domain scores and overall scores, and (b) how does one report valid and reliable overall scores and domain scores? The validity of the test structure is an important issue. Oftentimes the dimensional structure of test data does not match with the domain structure designed by the content expert (Reckase, 2009); however, many testing companies and state assessments report domain scores based on the domain structure–this is the approach used in the study here. To improve objective-level score estimates, several methods that use the correlations between the objectives have been proposed (e.g., Ackerman & Davey, 1991; Kahraman & Kamata, 2004; Mislevy, 1987; Mislevy & Sheehan, 1989; Wainer et al., 2001; Wang, Cheng, & Chen, 2004; Yen, 1987). However, in all of the examples, the issues of how to obtain a composite/overall score from those models and what the best relationship between the domain/objective scores and overall scores is, were not discussed. Sheng & Wikle (2008) compared the item parameter recovery and model fit between two hierarchical structure models. One of the models uses the higher-order IRT (HO-IRT) model approach (assuming that each domain score is a linear function of the overall score); the other uses the Hierarchical multidimensional IRT (MIRT) model approach (assuming that the overall score is a linear function of all the domain scores). The HO-IRT model assumes linear relationships between overall scores and objective/domain scores in a higher order; these relationships then determine the correlations between the domain scores. There are two major limitations to this approach: (a) the relations between overall scores and domain scores are all linear; and (b) HO-IRT is a multi-unidimensional model, which essentially assumes a unidimensional model, and would work better when the correlations between domains are high. The second MIRT approach, like the simple averaging method, assumes that the overall score is a regression onto the set of domain scores. The simulation condition used in Sheng & Wikle (2008) favors the MIRT model (i.e., MIRT is the true

model used to simulate responses), and only the item parameter recoveries were considered. The two hierarchical models in this study were based on different assumptions concerning the linear relationship between the overall ability and the domain abilities.

Factor analysis and MIRT have been studied by many researchers for more than 20 years (Muraki & Carlson, 1995; Reckase, 1985; Reckase & McKinley, 1991; Segall, 1996), with estimation software such as TESTFACT (Wilson & Gibbons, 1987) and NOHARM (Fraser & McDonald, 1988). With the rise of the Markov chain Monte Carlo (MCMC) algorithm (Béguin & Glas, 2001; Patz & Junker, 1999a, 1999b), item parameter estimates for complex models such as the MIRT partial credit model for both exploratory and confirmatory modes can be achieved (BMIRT, Yao, 2003; Mplus, Muthén, & Muthén, 2004). MIRT for domain scores that borrow information from each other is reliable and accurate (Yao & Boughton, 2007). In Yao & Schwartz (2006), multidimensional test information functions for tests of mixed format (with both dichotomous and polytomous items) were introduced, and the resulting composite score from two-dimensional subscores was briefly discussed; the composite score that has the smallest standard error of measurement compared with the composite score obtained by any other linear combinations of the domain scores can be obtained and is more desirable. There is no linear relation assumed between overall score and domain score.

Taking into consideration practical needs, the complexity of real-data structures, and currently available methods, this research used both real and simulation data to investigate how to obtain a set of reliable score reports, which include domain scores and an overall score. Four methods were applied: (a) the unidimensional IRT model (UM); (b) the HO-IRT model, to simultaneously estimate the overall ability and domain abilities; (c) the MIRT model, to estimate the domain abilities, applying the maximum information method to obtain the overall ability from those domain abilities; and (d) the bifactor general model approach (G), with the general dimension for the overall ability and the

domain specific dimensions for domain abilities. Estimates of overall ability, domain ability, and item parameters from the proposed methods were compared.

Methods

For all four estimation methods, BMIRT (Yao, 2003; Yao, Lewis, & Zhang, 2008) was used to obtain parameter estimates. The BMIRT software [1] uses a Metropolis-Hastings MCMC algorithm to estimate parameters for the four methods.

Multidimensional IRT Models

Following the notation of the MIRT model in Yao & Schwartz (2006), for a dichotomously-scored item j, the probability of a correct response to item j for an examinee with ability $\vec{\theta}_i = (\theta_{i1}, \cdots, \theta_{iD})$ for the multidimensional three-parameter logistic (M-3PL; Reckase, 1997) model is:

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

where $x_{ij} = 0$ or 1 is the response of examinee i to item j. $\vec{\beta}_{2j} = (\beta_{2j1}, \cdots, \beta_{2jD})$ is a vector of dimension D for item discrimination parameters. β_{1j} is the intercept or the difficulty parameter, β_{3j} is the lower asymptote or the guessing parameter, and, with $\odot$ representing the inner or dot product of two vectors, $\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$. The parameters for the jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j})$. For a polytomously-scored 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 two parameter partial credit model(M-2PPC)

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

where $x_{ij} = 0, \cdots, K_j - 1$ is the response of examinee i to item j. $\beta_{\delta_k j}$ for $k = 1, 2, \ldots, K_j$ are the threshold 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 $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta_2 j}, \ldots, \beta_{\delta_{K_j} j})$.

Multidimensional Test Information Function and Standard Error of Measurement

The test information function was obtained by following Yao & Schwartz (2006). For an item j, following the M-3PL model, the information function at $\vec{\theta}$ is

$$I_j(\vec{\theta}) = \frac{(P_{j1} - \beta_{3j})^2(1 - P_{j1})}{P_{j1}(1 - \beta_{3j})^2}\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}, \tag{3}$$

where P_{j1} is as Equation 1, but without i, to indicate the probability at ability $\vec{\theta}$. For an item j, following the M-2PPC model, the information function at $\vec{\theta}$ is $I_j(\vec{\theta}) = \sigma^2\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}$, where $\sigma^2 = \sum_{k=1}^{K_j}(k-1)^2P_{jk} - (\sum_{k=1}^{K_j}(k-1)P_{jk})^2$, and P_{jk} is as Equation 2, but without i, and with $\otimes$ representing the outer product of two vectors, $\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}$ is a D×D matrix, and its mth row and nth column element is the product of the mth and nth element of $\vec{\beta}_{2j}$. The test information at $\vec{\theta}$ is $I(\vec{\theta}) = \sum_{j=1}^{J} I_j(\vec{\theta})$. The variance $V(\vec{\theta})$ can be approximated by $I(\vec{\theta})^{-1}$. At direction $\alpha = (\alpha_1, \cdots, \alpha_D)$, the composite score is $\theta_\alpha = \sum_{l=1}^{D} \theta_l w_l$, with variance $V(\theta_\alpha) = \vec{w}V(\vec{\theta})\vec{w}^T$, and standard error $SEM(\theta_\alpha) = V(\theta_\alpha)^{1/2}$. Here $\vec{w} = (\cos^2\alpha_1, \cdots, \cos^2\alpha_D)$, $\sum_{l=1}^{D} w_l = 1$, and $\vec{w}^T$ is the vector transpose of $\vec{w}$. Let the direction $\alpha_{\vec{\theta}} = (\alpha_1, \cdots, \alpha_D)$ be the solution such that $\vec{w}V(\vec{\theta})\vec{w}^T$ has minimum value for all possible angles. Then θ_α, the composite score at $\vec{\theta}$ in the direction $\alpha_{\vec{\theta}}$ will have maximum information, and thus will be the most reliable composite score.

Higher-Order IRT

For this model, the first order is an IRT model, which describes the item performance for a given domain ability. The second order describes linear relations between domain abilities and overall abilities. The domain abilities are expressed as linear functions of the overall ability, expressed as $\theta_{il} = \lambda_l\theta_i + \eta_{il}$, where $-1 < \lambda_l < 1$ is the latent coefficient in regressing the lth domain ability onto the overall ability $\theta_i \sim N(0, 1)$. $\eta_{il} \sim N(0, 1 - \lambda_l^2)$ is the error term that is independent of other error terms. Given the

overall ability and regression coefficient, $\theta_{il} \mid (\theta_i, \lambda_l) \sim N(\lambda_l \theta_i, 1 - \lambda_l^2)$. The correlation between domain abilities θ_{ik} and θ_{il} is $\lambda_k \times \lambda_l$. Note that for this model, an item can only belong to one domain, i.e., the item is simple structured, and the MCMC sampling procedure is different from MIRT in the previous section; it samples the overall ability θ_i from a normal distribution, samples the regression coefficient, and then samples the domain abilities based on the overall ability and the regression coefficients (de la Torre & Hong, in press). The overall ability and its MCMC sampling can be found in the Appendix.

Bifactor General Model

This model is similar to MIRT, but with all the items loading on the general dimension and domain specific items loading on domain specific dimensions. The general dimension and the domain specific dimensions are orthogonal to each other, and provide overall ability and domain ability estimates, respectively. The main purpose of the general model is to measure the overall ability; the domain specific dimensions are nuisance dimensions–they account for the residual.

Real Data

Real data were used to investigate the performance of different methods by comparing overall score estimates and domain score estimates. Student data from the $TerraNova^{TM}$ test for the five main content areas: Reading (46 items), Language (34 items), Mathematics (57 items), Science (40 items), and Social Studies (40 items) were cleaned and merged with 3953 cases that had responses for all five content areas. The proposed four methods (UM, HO-IRT, MIRT, and G) were used to obtain the overall score and domain scores, with the following naming convention and explanations:

- M-1 or UM: Unidimensional IRT calibration to obtain unidimensional ability estimates as the overall score by combining all responses from the five content areas, with

the 217 items. The correlations between the 5 content areas are all greater than .68. With the 217 items being used, this overall score estimate was used as the "true" ability for the student and was compared with the overall ability estimated from the other methods.

- M-2: Unidimensional IRT calibration to obtain a unidimensional score report for each content area (i.e., domain score) from the response data for that content. Overall scores were obtained by averaging the five content scores.

- M-3: Confirmatory 5-dimensional IRT calibration to obtain the 5-dimensional domain ability estimates as the score report for the 5 content areas, combining all responses, with the 217 items. The dimensional loadings follow the content design, e.g., items for mathematics load on the mathematics dimension. This is the MIRT approach to obtain domain scores. The overall scores were obtained by two methods: (a) M-3-1: by averaging the domain scores from MIRT; and (b) M-3-2: by using the maximum information function from MIRT.

- M-4: HO-IRT to obtain overall scores and domain scores, using the 217 items. A linear relationship between the overall score and domain scores was assumed and the coefficients were estimated simultaneously.

- M-5 or M-G: Bifactor analysis with 6-dimensions. The first dimension was the general dimension that had all the 217 items loading on it and it gave estimates for the overall ability. The other 5 dimensions were content specific dimensions and they were orthogonal to each other and to the general dimension; the correlations between the dimensions were all 0.

For the real data, overall ability estimates from M-1 were used as the "true" values to compare the relative performances between different models. The decision was based on the current practice of using unidimensional IRT for students score estimates for many assessments, even though tests are often designed to measure a few subskills with a certain number of items in each subskill. Typically, the correlation between subskills ranges from

.3 to .8.

The graphic structure of four models is illustrated in Figure 1, with six items and two domains.

Insert Figure 1 about here

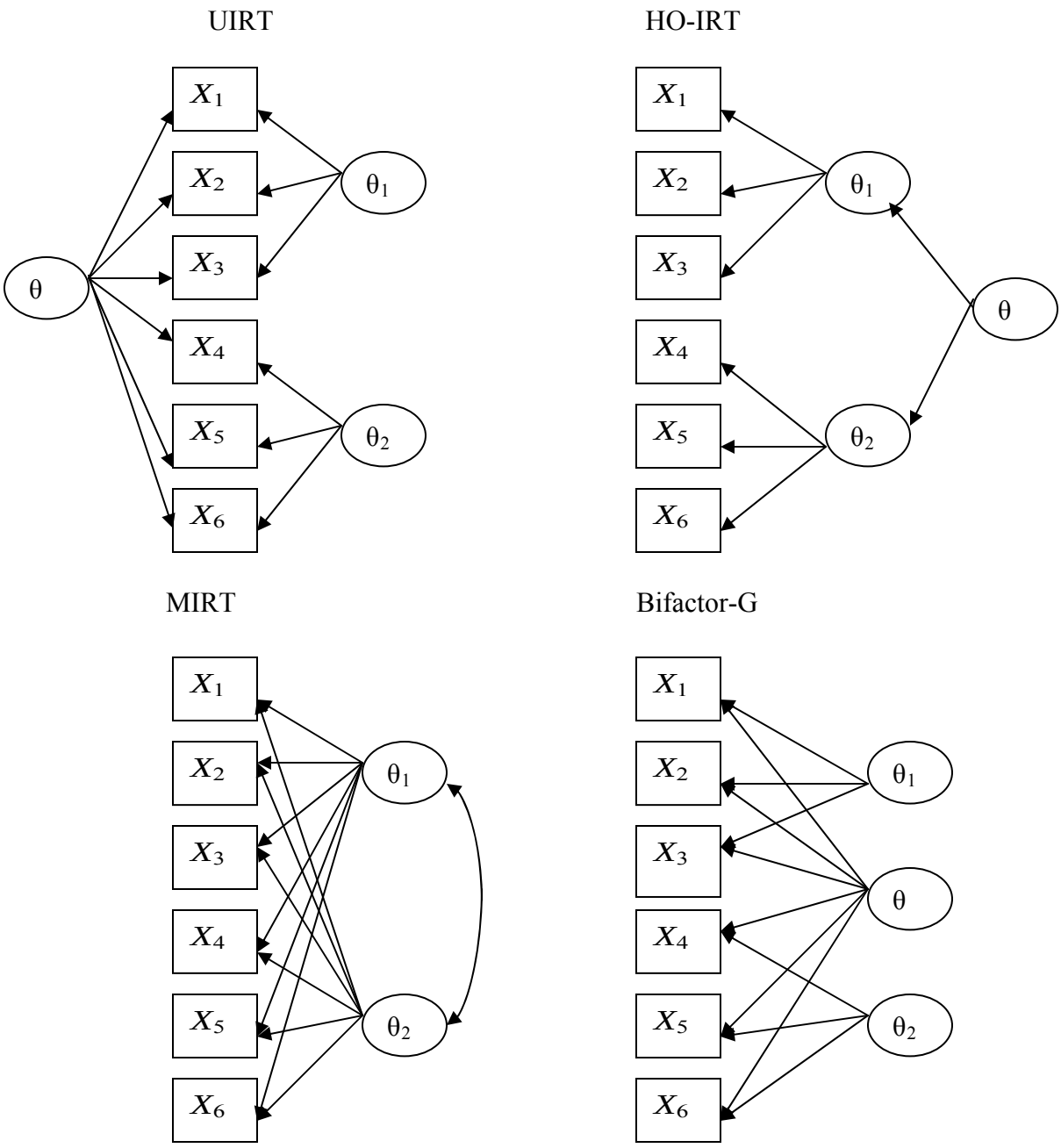

Figure 1. Graphical illustration for unidimensional model, higher-order model, multidimensional model, and bifactor model with six items and two domains. Note: Xj = Response to item j, θ_i = Domain ability i, θ = Overall ability.

Simulation Study

To evaluate the performance of the four methods, a simulation study was conducted. The conditions varied are listed in Table 1. First, sample sizes of $N = 500, 1000, 2000$ were used. Second, two types of models were used as the true model to simulate the data: HO-IRT and MIRT. Using the MIRT model as the true model, 4 domain abilities $(\theta_{i1}, \cdots, \theta_{i4})$, $i = 1, \cdots, N$ were sampled from standard multivariate normal distributions with correlations between domains of $r = 0, .3, .5, .7$, or .9. It is hypothesized that data with higher correlations between domains (essentially unidimensional) would favor HO-IRT, while data with lower correlations would favor MIRT (Yao, & Boughton, 2007). The true overall abilities were obtained from the domain abilities by the maximum information method. Using the HO-IRT model as the true model, the overall abilities θ_i, $i = 1, \cdots, N$, were sampled from $N(0, 1)$, and the domain abilities θ_{il} were obtained by sampling from $N(\lambda_l \theta_i, 1 - \lambda_l^2)$, $l = 1, \cdots, 4$. Three choices of $\lambda = (\lambda_1, \cdots, \lambda_4)$ for models refered to as HO-IRT-1, HO-IRT-2, and HO-IRT-3, were used. The first used $\lambda_1 = \cdots = \lambda_4 = .4$; the second used $\lambda_1 = \cdots = \lambda_4 = .9$; and the third used $\lambda_1 = .4$, $\lambda_2 = .7$, $\lambda_3 = .2$, $\lambda_4 = .9$. The three choices of λ gave different combinations of correlations between domain abilities. Third and last, four tests were designed that vary by test length. The items were selected from among 60 items that were obtained from a large-scale grade 8 mathematics assessment, with four domains of 15, 15, 12, and 18 items, respectively. Fifteen out of the total of 60 items were polytomously-scored items and had 3, 4, or 5 response categories. The calibration sample consisted of responses from 10,000 examinees to the 60 items, which were fit to a four-dimensional solution following the test blueprint. Three tests were selected: Test 1, Test 2, and Test3, with totals of 20, 32, and 48 items, respectively, with each test organized into 4 domains consisting of 5 (Test 1), 8 (Test 2), and 12 (Test 3) items within each of the four domains. For each domain, the chosen items represented the domain well. Test 4 contained all 60 items. Varying

simulation conditions were as follows: (a) eight true models; (b) four test designs; (c) three sample sizes; and (d) four estimation methods, with 20 replications for each condition. For each of the $8 \times 4 \times 3 \times 20 = 1920$ response data sets, the four methods (UM, HO-IRT, MIRT, and G) were used to obtain the overall score and domain scores. To compare the true score with the estimated score, the metric or scale of the estimated parameters must be the same as that of the true value. In estimating the item and ability parameters simultaneously for the MIRT model, one solution is to fix the population parameters (through priors) at their true values. These true values are unknown, but can be approximated by replacing them by the means and variance-covariance matrix obtained from raw (or response) data. This is the approach used in this study. However, this approach still cannot guarantee that the metrics of the estimates are the same as the true values. Therefore, domain abilities and overall abilities were also estimated by fixing the item parameters at their true values for some conditions.

Table 1

Conditions for the Simulation Study

True-Model	Population distribution	Sample size	Test
HO-IRT-1	$\lambda_1 = \cdots = \lambda_4 = .4$	500	Test1, J=20
HO-IRT-2	$\lambda_1 = \cdots = \lambda_4 = .9$	1000	Test2, J=32
HO-IRT-3	$\lambda_1 = .4$, $\lambda_2 = .7$,$\lambda_3 = .2$,$\lambda_4 = .9$	2000	Test3, J=48
MIRT-1	MV, r=.0		Test4, J=60
MIRT-2	MV, r=.3		
MIRT-3	MV, r=.5		
MIRT-4	MV, r=.7		
MIRT-5	MV, r=.9		

Evaluation Criteria

Root mean squared error (RMSE), absolute bias (ABS), bias (BIAS), and reliability were used to evaluate the accuracy of overall ability and domain ability parameter recoveries. RMSE and the test response function (TRF) were used to evaluate the item parameter recovery. They are defined as follows: let f_{true} be the value of a function obtained from the true parameter and f_l be the value of a function obtained from the estimated parameters from sample l. Here the function f can represent item discrimination, item difficulty, TRF, ability parameters, etc. RMSE was calculated by $RMSE = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(f_l - f_{true})^2}$, where n is the number of replications. ABS and BIAS were defined by $ABS = \frac{1}{n}\sum_{l=1}^{n} \mid f_l - f_{true} \mid$, $BIAS = \overline{f} - f_{true}$, where $\overline{f} = \frac{1}{n}\sum_{l=1}^{n} f_l$ is the final estimate. To evaluate the recovery of TRF, let f be the TRF function defined below:

$$TRF(\vec{\theta}, \vec{\beta}) = \frac{1}{J_1 + \sum_{j=1}^{J_2}(K_j - 1)} \sum_{i=1}^{N}[\sum_{j=1}^{J_1} P_{ij1} + \sum_{j=1}^{J_2}\sum_{k=1}^{K_j}(k-1)P_{ijk}], \tag{4}$$

where J_1 is the number of multiple choice items and J_2 is the number of constructed response items. Reliability was obtained by the following:
Reliability=$\frac{1}{n}\sum_{l=1}^{n} cor(f_l, f_{true})^2$, where f represents the overall ability or domain ability and *cor* is the correlation function. If the ability estimates recover the true value well, then higher correlation or reliability and lower RMSE or ABS should be observed. Finally, analysis of variance (ANOVA) for the RMSE was conducted to examine the effect of the varying conditions for the simulation study.

Results

Results for the Simulation Study

For each of the conditions (1920 response data sets), the four methods–UM, HO-IRT, MIRT, and G–were used through running BMIRT to obtain item parameter and ability parameter estimates simultaneously. Using response data for the 8 true models and 20 replications (total of 160 response data sets), for Test 4 and a sample size of 2000, BMIRT was also performed by fixing the item parameters at their true values to obtain the ability estimates for the HO-IRT and MIRT methods. I found that the ability recovery from the two methods (fixing item parameters versus not fixing) were quite similar. Therefore, all the estimates were comparable with the true values. Note that the purpose of both the G and the UM models was to obtain overall ability; therefore, the item parameter estimates from the two methods could not be compared with the true values.

Item Parameter Recovery. First, I examined the multidimensional item parameter recovery using HO-IRT and MIRT estimation methods for all the conditions. Table 2 displays the RMSEs for the item parameter estimates from sample sizes of 500 and 2000 for Test 4 (all items) and Table 3 shows the RMSEs for the TRF. The means for the true item parameters are listed in the first few rows. Note that there are only two threshold parameters being presented for constructed response items. Overall, HO-IRT

and MIRT recovered the item parameters similarly, with MIRT performing slightly better than HO-IRT for most of the conditions. The RMSEs for the item parameters decreased as the correlations between dimensions increased and as the sample size increased for all the estimation methods and true models. However, the RMSEs for the TRF increased slightly as the correlations increased.

Table 2

RMSE for the Item Parameter Estimates for Sample Size of 500 and 2000 for Test 4 for the Simulation Study

Sample Size	Item Type	True Model	Estimation-Method					
			HO-IRT			MIRT		
		3PL	β_{2j}	β_{1j}	β_{3j}	β_{2j}	β_{1j}	β_{3j}
		Mean	1.700	.910	.180			
		2PPC	β_{2j}	$\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$	β_{2j}	$\beta_{\delta_2 j}$	$\beta_{\delta_3 j}$
		Mean	1.172	-.508	1.149			
	MC	MIRT-1	.408	.406	.031	.405	.402	.030
		MIRT-2	.402	.408	.030	.400	.403	.030
		MIRT-3	.388	.414	.030	.388	.411	.030
		MIRT-4	.379	.419	.030	.373	.414	.030
		MIRT-5	.363	.433	.030	.357	.430	.030
500	CR	MIRT-1	.149	.178	.219	.149	.176	.218
		MIRT-2	.150	.181	.232	.150	.177	.224
		MIRT-3	.140	.182	.247	.141	.178	.241
		MIRT-4	.132	.192	.253	.130	.189	.250
		MIRT-5	.124	.203	.273	.122	.204	.270
	MC	HO-IRT-1	.406	.412	.030	.403	.407	.031
		HO-IRT-2	.390	.456	.030	.381	.453	.030
		HO-IRT-3	.383	.434	.031	.385	.433	.031
	CR	HO-IRT-1	.146	.196	.251	.149	.194	.249
		HO-IRT-2	.122	.234	.301	.121	.232	.297
		HO-IRT-3	.135	.211	.274	.132	.206	.272
	MC	MIRT-1	.233	.276	.025	.232	.277	.025
		MIRT-2	.227	.272	.025	.227	.274	.025
		MIRT-3	.227	.269	.025	.226	.270	.025
		MIRT-4	.225	.265	.024	.223	.266	.024
		MIRT-5	.217	.259	.024	.216	.258	.024
2000	CR	MIRT-1	.082	.089	.105	.081	.089	.104
		MIRT-2	.076	.091	.109	.076	.091	.109
		MIRT-3	.071	.095	.116	.072	.097	.117
		MIRT-4	.069	.098	.121	.069	.098	.120
		MIRT-5	.066	.098	.120	.064	.099	.118
	MC	HO-IRT-1	.238	.282	.025	.236	.280	.025
		HO-IRT-2	.226	.264	.024	.220	.262	.024
		HO-IRT-3	.217	.271	.025	.216	.271	.025
	CR	HO-IRT-1	.074	.091	.118	.073	.090	.117
		HO-IRT-2	.072	.102	.131	.068	.100	.126
		HO-IRT-3	.073	.093	.124	.073	.091	.121

Note: MC = multiple choice, CR = constructed response.

Table 3

RMSE for the TRF for Test 4 for the Simulation Study

True	500		1000		2000	
Model	HO-IRT	MIRT	HO-IRT	MIRT	HO-IRT	MIRT
MIRT-1	.0039	.0036	.0044	.0044	.0033	.0032
MIRT-2	.0072	.0062	.0062	.0059	.0043	.0042
MIRT-3	.0009	.0081	.0075	.0067	.0049	.0048
MIRT-4	.0105	.0099	.0083	.0072	.0054	.0050
MIRT-5	.0131	.0123	.0083	.0082	.0058	.0054
HO-IRT-1	.0069	.0064	.0048	.0042	.0032	.0031
HO-IRT-2	.0146	.0139	.0121	.0111	.0070	.0063
HO-IRT-3	.0108	.0104	.0088	.0081	.0051	.0048

Domain Score Recovery. Four domain ability estimates from HO-IRT and MIRT were compared with the true values through examinations of the reliability, RMSE, BIAS, and the classification rate. Expected results were observed. As the test length increased, as the correlation between dimensions increased, and as the sample size increased, the reliability increased and RMSE and BIAS decreased for all the models and methods. The MIRT estimation method performed slightly better than HO-IRT for all the criteria and for all the conditions, but the differences between the two methods were minor. In order to save space, the relative performance between the HO-IRT and MIRT methods are demonstrated by showing only the results for sample size 2000 and Test 4 in Table 4, because the other conditions gave similar results. As shown, the reliabilities for all the domain scores were higher than, or around, .8 for all the varying correlations or coefficients, except for θ_3 (which had the least number of items) when the true models had small correlations. Though it was not shown, the domain ability estimates from the G

model were compared to the true values, and it was found that as the correlation between dimensions increased, the reliability decreased and the RMSE increased. For example, for Test 4 and sample size 2000, the reliabilities went from .8 to .1 and the RMSEs went from .4 to .8 as the correlation increased from 0 to .9. Table 5 shows the classification rates for the 4 domain scores from HO-IRT, MIRT, and G methods for Test 4 and sample size 2000 for HO-IRT-1 and MIRT-4. As cut scores from the standard point of view are out of the scope of this study, a simple classification into three levels ("fail", "average", and "advanced") was used to compare different approaches, with ability below -α indicating "fail", above α indicating "advanced", and "average" otherwise. The definition of α is displayed at the bottom of the table. The first column indicates the level from the true data (A-"advanced"; V-"average"; F-"fail"), and the second column indicates the level based on estimates for the four domains, with their percentages indicated by the remaining columns. HO-IRT and MIRT yielded very similar results and the G model resulted in the worst match. MIRT performed better than HO-IRT at the two ends and worse than HO-IRT for mid-level abilities.

Table 4

Reliability, RMSE, and BIAS for the Domain Scores for Sample Size of 2000 and Test 4 for the Simulation Study

True Model	Estimation-Methods: HO-IRT				MIRT			
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
				Reliability				
MIRT1	.824	.812	.759	.818	.826	.813	.762	.820
MIRT2	.829	.819	.776	.826	.830	.820	.778	.828
MIRT3	.839	.832	.798	.839	.841	.833	.800	.840
MIRT4	.858	.854	.832	.859	.859	.855	.833	.860
MIRT5	.898	.897	.891	.900	.899	.898	.892	.901
HO-IRT-1	.819	.810	.773	.828	.820	.812	.776	.829
HO-IRT-2	.878	.872	.854	.880	.879	.872	.856	.881
HO-IRT-3	.817	.834	.774	.845	.819	.836	.776	.846
				RMSE				
MIRT1	.415	.421	.459	.406	.412	.419	.456	.404
MIRT2	.409	.413	.447	.399	.408	.412	.445	.397
MIRT3	.398	.399	.430	.387	.397	.398	.429	.386
MIRT4	.375	.375	.397	.366	.373	.373	.397	.365
MIRT5	.320	.318	.326	.313	.318	.316	.325	.312
HO-IRT-1	.406	.420	.459	.410	.404	.418	.457	.408
HO-IRT-2	.350	.353	.378	.350	.349	.351	.377	.349
HO-IRT-3	.406	.404	.459	.395	.404	.402	.457	.393
				BIAS				
MIRT1	-.003	-.042	-.016	-.041	-.001	-.042	-.016	-.041
MIRT2	-.013	-.047	-.032	-.052	-.013	-.048	-.031	-.053
MIRT3	-.021	-.051	-.041	-.056	-.022	-.053	-.041	-.058
MIRT4	-.029	-.053	-.047	-.057	-.028	-.052	-.045	-.056
MIRT5	-.039	-.052	-.050	-.055	-.037	-.051	-.048	-.053
HO-IRT-1	.009	-.033	-.011	-.074	.011	-.033	-.008	-.073
HO-IRT-2	-.065	-.059	-.060	-.077	-.061	-.055	-.056	-.073
HO-IRT-3	-.041	-.062	-.042	-.068	-.036	-.060	-.039	-.066

Table 5

Classification Rate for Domain Scores from Different Models for Test 4 and Sample Size of 2000 for the Simulation Study

Type			θ_1			θ_2			θ_3			θ_4		
			HO	G	MIRT	HO	G	MIRT	HO	G	MIRT	HO	G	MIRT
True-HO-IRT-1														
A	A		13.1	10.8	13.2	12.9	11.6	13	10.7	9.0	10.9	13.3	11.2	13.3
A	V	-	2.2	4.5	2.1	3.3	4.6	3.2	5.9	7.7	5.8	4.4	6.5	4.4
V	A	+	.1	.0	.1	.1	.0	.1	.2	.2	.2	.0	.3	.0
V	V		67.7	67.1	67.7	67.2	66.4	67.2	67.3	67.1	67.3	66.4	64.9	66.5
V	F	-	.3	.3	.2	.2	.4	.2	.2	.3	.2	.3	1.5	0.2
F	V	+	4.8	.35	4.9	5.0	8.0	5.1	4.2	6.8	4.0	3.3	5.9	3.4
F	F		12	8.4	11.9	11.5	8.5	11.5	11.7	9.1	11.9	12.4	9.8	12.4
True-MIRT-4														
A	A		13.1	3.3	13.2	13.0	3.5	13.2	11.6	.9	11.7	13.5	3.9	13.7
A	V	-	3.5	13.3	3.4	3.6	13.0	3.4	4.6	15.3	4.5	3.6	13.1	3.0
V	A	+	.2	.9	.2	.3	.6	.3	.3	.5	.3	.2	.9	.2
V	V		63.4	62.8	63.3	65.4	65.2	65.3	66	67.3	65.9	65.8	65.2	65.6
V	F	-	1.7	1.6	1.7	1.7	1.5	1.8	1.9	.5	2.0	1.7	1.4	1.9
F	V	+	4.3	17.7	4.2	3.4	15.7	3.5	3.0	15.2	2.9	2.9	15.2	2.9
F	F		14.0	.6	14.1	12.8	.5	12.8	12.7	.5	12.8	12.9	.6	12.9

Note: Cut Score $\alpha = 1$.

Overall Score Recovery. Overall ability estimates from UM, HO-IRT, MIRT with maximum information, and G were compared with the true value by examining reliability, RMSE, BIAS, and the classification rate. Only the reliabilities for sample size of 2000 are presented in Table 6 to demonstrate the results. Some results similar to the domain estimates were observed. For MIRT as the true model, the MIRT maximum information method gave the best results for all the criteria. For HO-IRT as the true model, HO-IRT gave the best results for reliability and RMSE when the correlation was low (HO-IRT-1, HO-IRT-3), followed by UM and G; the MIRT method performed as well as HO-IRT when the correlation was high (HO-IRT-2). Note that unlike the domain score, the reliability for the overall score was small when the correlation between dimensions was small for all the methods, especially for the HO-IRT and G methods; when the correlation was zero and the true model was MIRT, the HO-IRT method gave a reliability of less than

.2, while the MIRT maximum information method gave a reliability of .78 for Test 4, for example. When the true correlation between dimensions was about .7, the reliability was higher than .8 for all four tests and all four methods. For HO-IRT-2, the coefficient was $\lambda = .9$, which implied that the correlation between dimensions was .8, higher than the correlation in MIRT-4 (r=.7). However, the reliability for HO-IRT-2 was smaller than MIRT-4 for all four methods. Figure 2 shows 500 examinees' overall scores (x-axis) and their estimates (y-axis) from Test 4 and the four estimation methods: MIRT, HO-IRT, UM, and G. We can see that MIRT tended to overestimate more than HO-IRT did for lower ability levels. Larger BIASes were observed for all four methods when the true model was HO-IRT-1, HO-IRT-3, and MIRT-1. For MIRT-1, the correlation between dimensions was 0; therefore, the HO-IRT method performed very poorly, as the correlations in the HO-IRT method cannot be 0. Table 7 shows the classification rate for the overall scores from the HO-IRT, MIRT, UM, and G methods for Test 4 and sample size 2000 when the true model was HO-IRT-1 and MIRT-4. Similar to the analysis of domain ability, three levels were defined. Different cut scores were used for different simulation conditions or true models; true score ranges were different for different simulation conditions. For HO-IRT-1 as the true model, MIRT gave the worst match at the two ends and a better match for the mid-ability. The percentage of mismatch was relatively high for all four methods. It was also found that when the true model was MIRT-1, all methods yielded poor matches, with the exception of the MIRT method.

Table 6

Reliability for Overall Scores from Test 1-4 for Sample Size of 2000 for the Simulation Study

Test	True Model	Estimation-Methods HO-IRT	MIRT	UM	G
	MIRT1	.163	.580	.329	.174
	MIRT2	.696	.724	.692	.575
Test 1	MIRT3	.767	.782	.767	.749
	MIRT4	.815	.823	.818	.816
	MIRT5	.849	.854	.854	.854
	HO-IRT-1	.335	.315	.292	.204
	HO-IRT-2	.800	.801	.804	.804
	HO-IRT-3	.626	.509	.609	.621
	MIRT1	.183	.697	.362	.261
	MIRT2	.766	.809	.757	.561
Test 2	MIRT3	.828	.849	.830	.789
	MIRT4	.870	.878	.872	.866
	MIRT5	.896	.899	.899	.898
	HO-IRT-1	.373	.346	.289	.186
	HO-IRT-2	.842	.843	.843	.841
	HO-IRT-3	.668	.518	.616	.581
	MIRT1	.157	.744	.403	.465
	MIRT2	.795	.844	.805	.779
Test 3	MIRT3	.855	.878	.859	.851
	MIRT4	.896	.902	.897	.895
	MIRT5	.919	.921	.921	.921
	HO-IRT-1	.396	.370	.376	.361
	HO-IRT-2	.867	.866	.867	.868
	HO-IRT-3	.704	.533	.657	.700
	MIRT1	.188	.786	.367	.469
	MIRT2	.816	.874	.830	.795
Test4	MIRT3	.877	.904	.883	.872
	MIRT4	.915	.924	.917	.914
	MIRT5	.937	.939	.938	.938
	HO-IRT-1	.412	.376	.340	.366
	HO-IRT-2	.884	.883	.884	.883
	HO-IRT-3	.748	.606	.713	.741

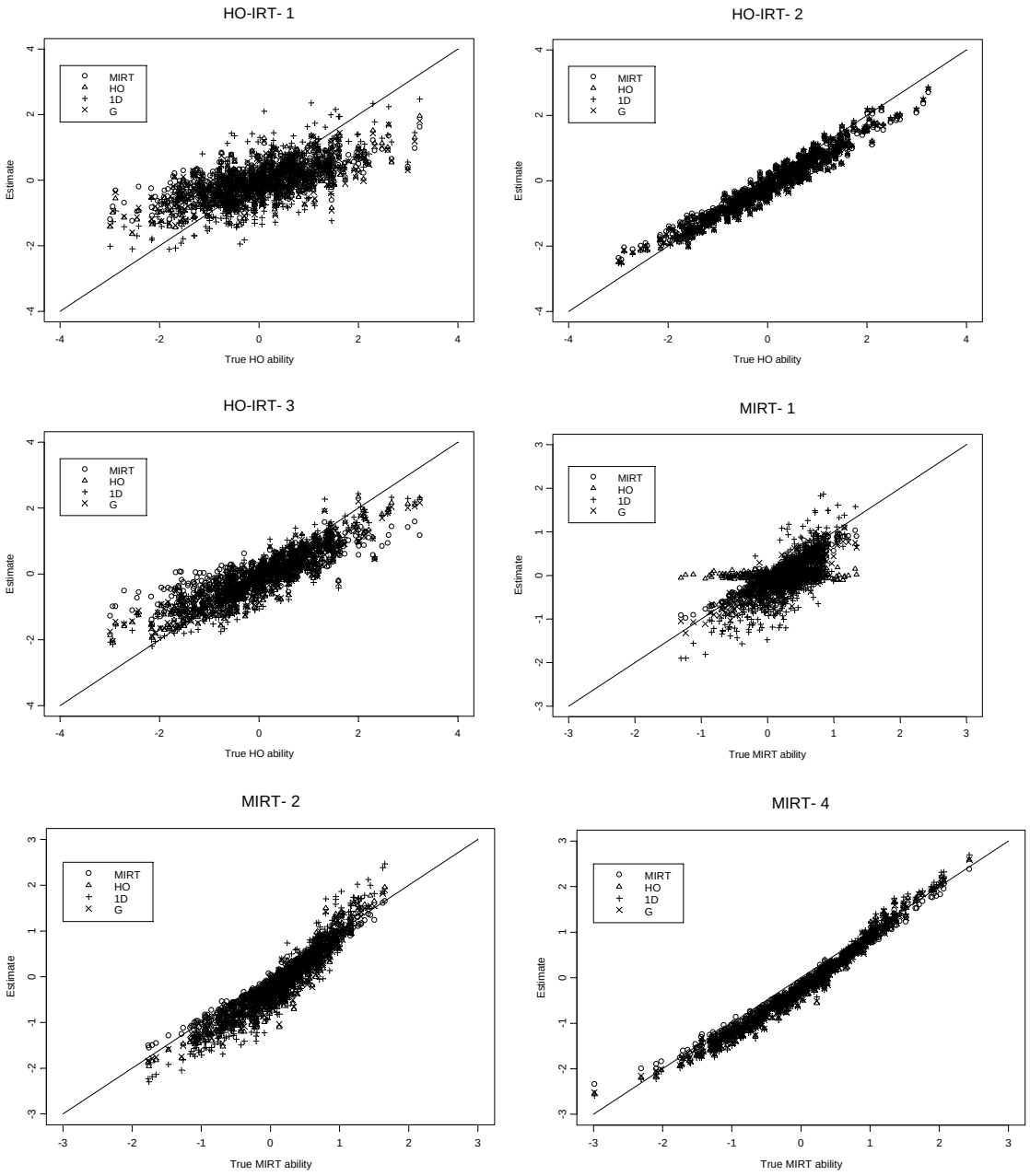

Figure 2. Comparison of the 500 true overall scores with the estimates from different models for test 4 for the simulation study.

Table 7

Classification Rate for Overall Ability from Different Models for Test 4 and Sample Size of 2000 for the Simulation Study

Type			HO-IRT-1				MIRT-5			
			HO-IRT	G	UM	MIRT	HO-IRT	G	UM	MIRT
A	A		3.2	3.0	6.6	2.3	13.8	13.5	13.9	13.2
A	V	-	13.5	13.7	9.9	14.4	1.6	1.8	1.6	2.2
A	F	-	.0	.0	.2	.0	.0	.0	.0	.0
V	A	+	1.1	1.0	5.5	.9	0	.1	.05	.0
V	V		65.6	65.9	56	66.6	67	67.2	66.9	67.9
V	F	-	.8	.6	6.0	.1	1.8	1.5	1.8	.9
F	A	+	.0	.0	.1	.0	.0	.0	.0	.0
F	V	+	13.5	13.7	9.6	15.2	.2	.2	.2	.4
F	F		2.4	2.2	6.2	.8	15.8	15.7	15.8	15.6

Note: Cut Score $\alpha = 1$.

Relations between Overall Score and Each of the Domain Scores. Table 8 shows the linear regression coefficient for the true value, or obtained from the true value, and those estimated from the MIRT and HO-IRT methods for Test 4 and sample size 2000 obtained by the formula: $\theta = \lambda_1\theta_1 + \lambda_2\theta_2 + \lambda_3\theta_3 + \lambda_4\theta_4$. When the true model was HO-IRT, the coefficients estimated from HO-IRT were closer to the true values; when the true model was MIRT, the coefficients estimated from MIRT were closer to the true values.

Table 8

Coefficient of Linear Regression between the Overall Score and the Domain Scores for Sample Size of 2000 and Test 4 for the Simulation Study

True	Estimation-Methods											
Model	True				MIRT				HO-IRT			
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
MIRT1	.18	.19	.15	.17	.23	.21	.14	.23	-.02	.04	.03	.03
MIRT2	.20	.20	.17	.19	.25	.22	.15	.24	.26	.27	.33	.17
MIRT3	.22	.23	.18	.20	.26	.23	.16	.25	.26	.27	.31	.26
MIRT4	.24	.25	.19	.21	.26	.24	.18	.26	.25	.26	.29	.25
MIRT5	.26	.26	.21	.22	.27	.25	.19	.27	.24	.24	.28	.24
HO-IRT-1	.29	.28	.32	.26	.23	.21	.15	.24	.25	.28	.30	.24
HO-IRT-2	.26	.24	.27	.26	.26	.24	.18	.27	.27	.24	.26	.25
HO-IRT-3	.06	.20	.02	.75	.23	.22	.16	.24	.07	.23	.04	.70

Correlations between Domain Scores. Correlations between domain score estimates from MIRT and HO-IRT were compared with the true values. We found that both methods gave higher correlations than the true values, with HO-IRT giving a slightly higher correlation than MIRT. In MCMC ability sampling, because of the effect of priors, correlated subscale scores resulted in correlated errors, and ultimately, the correlations between the estimated abilities were higher than the true values. Tables for the correlations are not presented here. The range of the increase in this study was .0 to .15.

Analysis of Variance. For the RMSEs of the overall ability, an ANOVA procedure was conducted to examine the effect of varying factors: estimation method, test length, population or correlation between dimensions, and sample size. Tukey's Studentized Range (HSD) Test for the RMSEs showed that the UM and G methods did not differ from

each other, but they were significantly different from HO-IRT and MIRT; HO-IRT and MIRT were significantly different from each other. Both correlations between dimensions and test length had significant effects, except that (a) MIRT-4 (r=.7) and MIRT-5 (r=.9) were not different; and (b) Test 3 and Test 4 were not different. Sample size did not have a significant effect (at α=.05). The ANOVA also confirmed that when the true model was HO-IRT, the HO-IRT method gave better results than MIRT, which in turn gave results better than G and UM. When the true model was MIRT, the MIRT method gave significantly better results than HO-IRT, which itself gave better results than G and UM. The four estimation methods were significantly different for all the true models except for HO-IRT-2 and MIRT-5, when the correlation between dimensions was higher than .8.

Results for Real Data Application

For the real data, four models M-1, M-3, M-4, and M-5, using unidimensional, MIRT, HO-IRT, and General, were applied to the combined data of 217 items. Model fit statistics are displayed at Table 9. The chi-square statistic is 2 times the difference of the loglikelihood function between M-1 and the other models. The last column is the total number of parameters for both the examinees and the items. Since we do not know the true model, only the relative performances of the models can be examined; the general model fit the data best, followed by MIRT, HO-IRT, and M-1. The loglikelihood function from MIRT was 1650 higher than those from HO-IRT, indicating that MIRT favored the data more than HO-IRT; since HO-IRT had more parameters to be estimated than MIRT did.

Table 9

Model Fit Statistics for the Real Data

Methods	LogLikelihood	Chi-Square	Df
M-1	-413598		4604=3953+651
M-3(MIRT)	-398466	30264	20416=19765+651
M-4(HO-IRT)	-400116	26964	24374=23718+651+5
M-5(M-G)	-397603	31990	24586=23718+868

Comparison of Overall Score Estimates. Figure 3 compares the overall ability estimates from different models with the "true" overall ability on the x-axis, estimated from UM (M-1; 217 items), and standard errors of measurement (SEM) for UM on the y-axis. The y-axis also shows the differences between the "true" overall abilities and overall ability estimates from models M-2 (simple average), M-3 (MIRT), M-4 (HO-IRT), and M-5, respectively.

Overall ability obtained from averaging the domain abilities (M-2) had the largest difference from UM; the differences were also larger than SEM at both ends. Estimates from the MIRT maximum information method (M-3-2) were lower than estimates from UM for low-ability examinees and slightly higher for high-ability examinees; estimates were better than the MIRT model with simple averaging (M-3-1). The HO-IRT (M-4) and G model were very similar, with G estimates showing a larger difference from UM estimates for more examinees. Table 10 shows the correlations, ABSes, and BIASes between the "true" overall abilities and their estimates from M-2, M-3-1, M-3-2, M-4, and G, averaged over all examinees. The correlations, ABSes, and BIASes between UM (M-1) and MIRT (M-3-2), UM and HO-IRT, and UM and G were similar. The ABS and BIAS between UM and M-2 (simple averaging) were the largest.

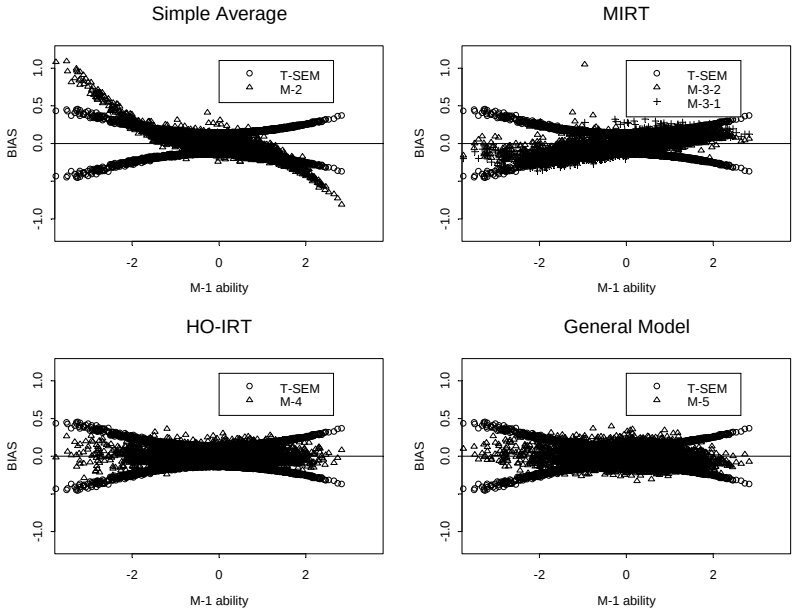

Figure 3. Comparison of overall score estimates from different models for TerraNova™ Data.

For the overall ability from M-1, the test information had a maximum value at about $\theta = 0$. For a standard normal distribution, about 72% of θ are between (-1, 1), and 14 percent of θs at each tail of the distribution are beyond these values. Simple classification of three levels ("fail", "average", and "advanced") was used to compare different approaches, with overall ability below -1 indicating "fail", above 1 indicating "advanced", and "average" otherwise. Table 11 shows the classification rates between different models, compared with the classification rate with M-1. The first column indicates the level (A-"advanced"; V-"average"; F-"fail") from M-1, and the second column indicates the level from other methods, with their percentage indicated by the remaining of the columns. Simple averaging gave the worst match compared with M-1 (fourth and fifth columns). The MIRT maximum information method M-3-2 gave a slightly better match on the two ends, but a worse match for the "average" level when compared with the HO-IRT and G model. M-3-2 classified 1.7% of students to be "fail", while M-1 classified them as "average"; HO-IRT and G classified 0.6% of students to be "fail", while M-1 classified them as "average". Note that both M-1 and HO-IRT essentially assumed that the combined data with the 217 items fit the unidimensional model, and the G model loaded all items on the general dimension.

Table 10

Comparison of the "True" Overall Ability (M-1) with

Estimates from Different Models for the Real Data

Methods	Corr	ABS	BIAS
M-2	.996	.120	.038
M-3-1	.998	.100	.017
M-3-2	.999	.057	-.024
M-4	.997	.058	.008
M-G	.995	.081	-.006

Table 11

Classification of the Overall Score Estimates from M-2, M-3-1, M-3-2, M-4, and M-5, Compared with those from the "True" Overall Ability

Type			M-2	M-3-1	M-3-2	M-4	M-5
A	A		11.9	16.3	16.1	15.7	15.0
A	V	-	4.4	3.0	.1	.6	1.4
V	A	+	.0	.0	1.1	1.0	.9
V	V		68.7	62.7	66.0	67.2	67.3
V	F	-	.0	3.1	1.7	.6	.6
F	V	+	1.2	.0	.3	.9	1.0
F	F		13.7	15.0	14.6	14.1	13.9

Note: Cut score α=1.

Comparison of Domain Score Estimates. Table 12 shows the results comparing the domain abilities from the unidimensional model (M-2) with those from MIRT (M-3), HO-IRT (M-4), and G, averaged over all examinees. The correlations for all the five content scores between the traditionally used method (M-2) and MIRT were slightly higher than those between M-2 and HO-IRT. The BIASes between M-2 and MIRT were slightly less than those between M-2 and HO-IRT; the ABSes between M-2 and MIRT were slightly bigger than those between M-2 and HO-IRT. MIRT overestimates higher ability examines and underestimates lower ability examinees. This might be caused by the fact that the domain abilities in MIRT borrowed information from each other, resulting in higher reliabilities and larger variances. The domain abilities from the general model were not as accurate as the other models. One would expect that the higher the correlation is between the dimensions, the more accurate the general dimension and the less accurate the other dimension is. This was also demonstrated by the simulation results.

Table 12

Comparison of Domain Ability Estimates from UM with Estimates from MIRT, HO-IRT, and G (General) Models for the Real Data

Content	M-3			M-4			M-5		
	Corr	ABS	BIAS	Corr	ABS	BIAS	Corr	ABS	BIAS
MA	.992	.168	.015	.989	.111	.015	.575	.672	-.015
RD	.991	.190	.020	.979	.151	.045	.386	.761	-.008
LA	.988	.196	.025	.967	.192	.049	.400	.755	-.004
SC	.981	.225	.023	.965	.206	.045	.530	.653	-.008
SS	.987	.204	.020	.976	.168	.042	.485	.700	-.013

Relations between Overall Score and Domain Score. For the real data analysis, estimates of the correlations (between domains) from HO-IRT were higher (ranges .87-.95) than other methods' estimates. By running linear regressions between the overall ability and the domain abilities, we obtained the following relationship: For HO-IRT, $\theta = .09 \times \theta_1 + .29 \times \theta_2 + .28 \times \theta_3 + .16 \times \theta_4 + .21 \times \theta_5$; for the MIRT maximum information method, $\theta = .25 \times \theta_1 + .24 \times \theta_2 + .16 \times \theta_3 + .13 \times \theta_4 + .17 \times \theta_5$. One can see that the relations obtained from the two methods were quite different regarding the content area of Mathematics (θ_1). To further examine the differences between the performance of MIRT and HO-IRT, I pulled out domain scores and the overall score for one extreme examinee: the five domain scores from M-2, MIRT, HO-IRT, and G were (.36, -2.8, -2.7, 1.0, .9), (.24, -2.5, -2.5, .7, .6), (.1 , -1.6, -1.6 ,.0, .1), and (1.9, -2.7, -2.5, 2.5, 2.7) respectively. Clearly, domain scores from MIRT were much closer to M-2 than HO-IRT and G. On the other hand, the overall scores from M-1, M-3-2, HO-IRT, and G were -1.0, -.2, -.9, and -.9, respectively. The overall scores from HO-IRT and G were closer to M-1 than the MIRT maximum information method.

Conclusions and Discussion

For overall scores, emphasis is on reporting the students' overall achievement or proficiency in answering all the test items, while domain scores focus on offering teachers and students more diagnostic information in each domain. A solution that can satisfy both needs is in high demand.

In the simulation study I investigated the performances of four methods in obtaining both the overall score and domain scores for test of mixed item types. I found that both HO-IRT and MIRT resulted in similar item parameter and domain ability recovery, with MIRT performing slightly better than HO-IRT for all the conditions. The MIRT maximum information method and HO-IRT outperformed UM and G with regards to overall ability recovery when there were low or zero correlations between domains; when the correlation was higher than .8, all four methods performed similarly. In order to use either the MIRT or HO-IRT method to report both domain scores and overall scores with greater reliability than .8, the correlation between domains has to be higher than .7 or the coefficient between the overall score and the domain score has to be higher than .9 for a test of 5 items for each domain. An increase in the test length will allow lower correlations between domains. The MIRT method is preferred over HO-IRT, particularly when the correlation between domains is low. Though model fit statistics for the simulated data were not shown, the MIRT model was found to be the best for all the simulation conditions judging by AIC.

The relationship between overall ability and domain ability is quite complex; representing it as a linear relationship may be inadequate. I found from real data analysis that the HO-IRT method, the MIRT maximum information method, and the Bifactor general method gave similar overall ability estimates when compared to the UM model; simply averaging domain abilities obtained from unidimensional IRT resulted in larger bias compared to UM. Domain score estimates from HO-IRT and MIRT were quite

similar, with MIRT having higher correlations but larger ABS than HO-IRT, when compared with the "true"(i.e., estimating from the UM). The two methods yielded different relationships between the overall score and domain scores. In addition, classification of students differed between the two methods in about 2% of cases. Since we do not know the true model for the real data, caution is advised when considering which model to use. The simulation study results can help with the decision; correlations between dimensions, model fit statistics from different models, and content-expert opinion should also be considered when choosing which method to use.

For the HO-IRT method, an item can only contribute to one domain (simple structure); for both real and simulated data, simple structure was assumed. However, in practice, some items may measure more than two domains (complex structure). For example, a math story problem will require some level of reading ability. For such data, the MIRT method is the best solution (see Figure 1). The priors and proposal functions in BMIRT all affect the estimation accuracy. Adjusting the parameters of those functions or proposing different distribution functions may improve the estimation accuracy to varying degrees for different methods. HO-IRT is a multi-unidimensional model, which essentially assumes a unidimensional model and a linear relationship between the overall score and the domain score. In contrast, the MIRT maximum information method for deriving an overall score does not assume any linear relation between the domain score and the overall score and takes into account the fact that their relationship may be different at different score points or ability levels. Moreover, the obtained overall score at each of the score points has the smallest estimation error and the maximum information. An innovation in the measurement history of score reporting is using the IRT based score, which takes into account item difficulties, in place of the classical number correct score, which treated all items as of equal difficulty. Using the MIRT maximum information method for the overall score instead of the traditionally-used simple averaging method, or any other linear

relationship method, is an analogous innovation.

The MIRT model has been widely studied in recent years. There are still no definite answers to the question of how to find the best dimensional structure and the best number of dimensions for a set of response data; research on this topic is greatly needed. In practice, domain scores are based on the validity of the test design. As I have shown here, MIRT can report a set of reliable overall scores and domain scores. With MIRT equating or linking (Yao, in press; Yao & Boughton, 2009), these scores can be compared across forms, samples, and years.

Appendix: HO-IRT Sampling

The domain abilities are expressed as linear functions of the overall ability, $\theta_{il} = \lambda_l \theta_i + \eta_{il}$. After some notation changes and dropped i, we obtain $\mathbf{Y} = \mathbf{X}\theta + \eta$, where $\mathbf{Y}^T = (\frac{1}{\sqrt{1-\lambda_1^2}}\theta_1, \cdots, \frac{1}{\sqrt{1-\lambda_D^2}}\theta_D)$, $\mathbf{X}^T = (\frac{\lambda_1}{\sqrt{1-\lambda_1^2}}, \cdots, \frac{\lambda_D}{\sqrt{1-\lambda_D^2}})$, $\eta = (\eta_1, \cdots, \eta_D)$, and $\eta_l \sim N(0, \sigma^2)$, for $l = 1, \cdots, D$. σ^2 is close to 1. Suppose the prior $\theta \sim N(0, 1)$, then the posterior distribution of the overall ability (Hastie, Tibshirani, & Friedman, 2001) is $\theta \sim N(c\sum_{l=1}^{D} \frac{\lambda_l \theta_l}{1-\lambda_l^2}, c)$, and $c^{-1} = 1 + \sum_{l=1}^{D} \frac{\lambda_l^2}{1-\lambda_l^2}$.

References

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

CTB/McGraw-Hill (2001). TerraNova Second Edition, Monterey, CA: Author.

CTB/McGraw-Hill (2008). TerraNova Third Edition, Monterey, CA: Author.

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

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

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267-269.

Hastie,T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. New York, NY: Springer.

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. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661-679.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73-90.

Muthén, L. K., & Muthén, B.O. (2004). Mplus User's Guide, version 3. Los Angeles, CA: Author.

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

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

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

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.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.

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

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement. 68*, 413-430.

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.

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.

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. (2004). *LinkMIRT: Linking of multivariate item response model.* [Computer software]. Monterey, CA: Defense Manpower Data Center.

Yao, L. (2010, in press). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement.*

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

Yao, L., & Bough, K. A. (2009). Multidimensional linking for tests containing polytomous items. *Journal of Educational Measurement. 46*, 177-197.

Yao, L, Lewis, D., & Zhang, L. (2008, April). *An introduction to the application of BMIRT: Bayesian multivariate item response theory software.* Training secession presented at the meeting of the National Council on Measurement in Education, New York: NY.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement. 30*, 469-492.

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

Acknowledgments

The views expressed are those of the author and not necessarrliy those of the Department of Defense, or the United States government.

Footnotes

[1]BMIRT produces item and ability estimates and model fit statistics for multi-dimensional and multi-group models in both exploratory and confirmatory mode. It is available from the author.